https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions

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)

Abstract. Progress in machine learning in conjunction with the increasing availability of relevant Earth observation data streams may help to overcome uncertainties of global hydrological models due to the complexity of the processes, diversity, and heterogeneity of the land surface and subsurface, as well as scale-dependency of processes and parameters. In this study, we exemplify a hybrid approach to global hydrological modeling that exploits the data-adaptiveness of machine learning for representing uncertain processes within a model structure based on physical principles like mass conservation. Our H2M model simulates the dynamics of snow, soil moisture, and groundwater pools globally at 1° spatial resolution and daily time step where simulated water fluxes depend on an embedded recurrent neural network. We trained the model simultaneously against observational products of terrestrial water storage variations (TWS), runoff, evapotranspiration, and snow water equivalent with a multi-task learning approach.

We find that 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 global hydrological models. The neural network learned hydrological responses of evapotranspiration and runoff generation to antecedent soil moisture state that are qualitatively consistent with our understanding and theory. Simulated contributions of groundwater, soil moisture and snowpack variability to TWS variations are plausible and within the large range of traditional GHMs. H2M indicates a somewhat stronger role of soil moisture for TWS variations in transitional and tropical regions compared to GHMs.

Overall, we present a proof of concept for global hybrid hydrological modeling in providing a new, complementary, and data-driven perspective on global water cycle variations. With further increasing Earth observations hybrid modeling has large potential to advance our capability to monitor and understand the Earth system by facilitating a data-adaptive, yet physically consistent, joint interpretation of heterogeneous data streams.

1 Introduction

Physically-based hydrological modeling is an essential tool to understand, monitor, and forecast the water cycle with an array of societal implications (Jiménez Cisneros et al., 2014). Still, global hydrological and land-surface models face many problems related to process representations and parameterizations, resulting in large uncertainties (Schellekens et al., 2017). State-of-the-art global hydrological models (GHMs) still largely disagree across all spatial and temporal scales due to challenges such as limited, biased, and uncertain data, the heterogeneity of processes, or a lack of process understanding (Haddeland et al.,

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions
Open Access
EGU

2011; Beck et al., 2017). While global water cycle observations are accumulating rapidly, a thorough integration with global hydrological modeling to overcome uncertainties is rarely facilitated due to the complexity and computational expenses.

In our data-rich era, different pathways have been proposed to utilize additional Earth observation data in hydrological modeling. Physically-based models can benefit from using spatially explicit parameters, which can be retrieved from Earth observation data. It is, for example, common to use spatio-temporally varying leaf area index as model parameter (e.g., Van Der Knijff et al., 2010) to account for vegetation dynamics. Furthermore, upscaling of local parameters to global scale—such as catchment parameters (Beck et al., 2016) or soil properties (Hengl et al., 2017)—can improve model performance. Further, 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. 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 system conditions (Sun et al., 2016) using additional data, 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, the use of (purely data-driven) deep learning approaches has been proposed and applied recently in hydrology, and experimental methods for (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, so-called hybrid modeling (Reichstein et al., 2019). Hybrid modeling aims at harvesting the information in Earth observation data efficiently by replacing uncertain processes with a machine learning model, while still maintaining model interpretability and physical consistency to a certain degree. Instead of exclusively relying on explicit process representations, hybrid models can learn from data in a flexible way, making use of the large amounts of Earth observations available. Hybrid modeling could help to advance the predictability, describability and understandability of land surface processes by dealing with some of these issues: replacing processes that are not well understood or hard to parameterize with a machine learning model can reduce model biases and increase the local adaptivity. Furthermore, the approach facilitates the incorporation of information from multiple data sources, which is a bottleneck in global hydrological models.

The applicability of hybrid modeling to global scale environmental modeling has been shown in Kraft et al. (2020), where a dynamic neural network has been used to parameterize a simple hydrological model that represents the major hydrological states of groundwater, soil moisture, and snow. The neural network was physically constrained using hydrological balance equations and optimized on the observation-based products of terrestrial water storage (TWS), snow water equivalent (SWE), evapotranspiration (ET), and runoff (Q). The data-driven assessment of hydrological states and fluxes allowed to circumvent some of the shortcomings of process-based modeling and gives a new perspective on the water cycle.

One of the key hydrological data products for diagnosing and understanding global land water cycle variations is the terrestrial water storage (TWS). TWS is an observation-based rasterized product that integrates the total of all water storages 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) but also in data-driven studies (Humphrey et al., 2016; Andrew et al., 2017; Rodell et al.,

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

2018). A consistent attribution of TWS variations to its components (like groundwater, snow, or soil moisture) is still outstanding as current model simulations do not produce consistent patterns due to uncertainties in the model structure and process description, forcing data, and parameter values (Güntner, 2008). Such an attribution is not trivial, especially as contiguous observations of these components are not available separately on global scale (e.g., groundwater) or limited (e.g., soil moisture, where satellite observations are only sensitive to the top soil layers). Thus, decomposition of TWS is either done locally using in situ data (e.g., Swenson et al., 2008), using large-scale hydrological modeling, which allows a global perspective, or with data-driven approaches (Andrew et al., 2017) that lack physical consistency.

In this study, we evaluate the potential of hybrid modeling for providing a complementary and data-driven perspective on the global water cycle variability based on carefully designed cross-validation analysis. We further develop the model proposed by Kraft et al. (2020) with some adjustments for improved robustness and physical consistency. Section 2 describes the datasets used, the hybrid hydrological model (H2M), and the model training and evaluation approach. Furthermore, we introduce a set of GHM simulations from the earth2Observe ensemble that were used as a reference to assess the performance (Sect. 3.1) and plausibility of the hybrid model simulations. Section 3.2 investigates the data-driven estimates of the hydrological responses, followed by Sect. 3.3, where the TWS decomposition from the different models are contrasted. In Sect. 4.1, the model performance is discussed in the context of the GHM models, followed by an assessment of the interpretability of the hydrological responses in Sect. 4.2. In Sect. 4.3, a more general assessment of the challenges and opportunities of the hybrid approach is provided.

2 Data and methods

2.1 Datasets

2.1.1 Meteorological forcing

Three time-varying meteorological datasets were used to force the model (Tab. 1). **i)** Precipitation observations were obtained from the Global Precipitation Climatology Project dataset (GPCP-1DD) v1.2 (Huffman et al., 2012). **ii)** Net radiation is provided by the SYN1deg Ed3A product (Doelling, 2017) of the Clouds and the Earth's Radiant Energy Systems (CERES) program (Wielicki et al., 1996). **iii)** We used air temperature from the CRUNCEP v8 dataset, a 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).

2.1.2 Static variables

A set of static variables was used to represent surface and subsurface conditions (Tab. 1). **i)** Soil properties from the soilgrids dataset (Hengl et al., 2017): *absolute depth to bedrock* and the average content across all soil layers of *bulk density*, *coarse fragments*, *clay*, *silt*, and *sand*. **ii)** Land cover fractions were calculated 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*,

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions

Table 1. Dataset overview: water cycle constraints, meteorological forcing and static variables with their native and aggregated spatial resolution, as well as their temporal resolution. The mathematical notation uses upper case for state variables and lower case for fluxes.

	Acr.	Math. notation	Spatial resolution		Temporal resolution	Dataset	Resources
Water cycle constraints			Native	Agg.			
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)
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_{n}	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°	-	Soilgrids 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)

Arc.=acronym, Agg.=aggregated

shrublands, and *forests*. **iii)** A digital elevation model was obtained from GTOPO30 (DOI/USGS/EROS, 1997). **iv)** In addition, fractions of groundwater-driven wetlands, regularly flooded wetlands, and the intersection of the them (Tootchi et al., 2019) were used.

The total of 22 static variables were spatially aggregated from their finer resolution to 1/30° to keep sub-cell variations, yielding a block of 30 latitude cells times 30 longitude cells times 22 variables, i.e., a total of 19 800 values per 1° cell. To reduce the dimensionality prior to feeding the data into the model, a simple convolutional autoencoder, consisting of an encoder, a bottleneck layer and a decoder, was used: The decoder consist of a stack of consecutively smaller convolutional neural network (CNN) layers that reduces the input block to a vector of size 30, the bottleneck layer. This process is then reverted in the decoder model, mapping the vector back to the input data. The model tries to reconstruct the input data but is forced to find a low-dimensional representation by the bottleneck (e.g., Goodfellow et al., 2016). The compressed dataset is used as model input and contains highly non-linear spatial features of the original data.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

2.1.3 Observational constraints

Four observational water cycle components were used to constrain the model. The datasets were aggregated to a spatial resolution of 1° (Tab. 1). Due to a varying temporal coverage, the common period of January 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 total terrestrial water storages. These include groundwater, soil moisture, surface water, biosphere-bound water, snow, and ice. Due to outliers in the dataset, observations below -500 and above 500 mm were removed. **ii)** Monthly ET estimates were retrieved from the global FLUXCOM-RS product (Tramontana et al., 2016; Jung et al., 2019), based on machine-learning driven upscaling from site-level FLUXNET eddy covariance measurements (Baldocchi et al., 2001) to global scale. 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 are available from the GRUN v1 dataset (Ghiggi et al., 2019). The procuct is based on an upscaling approach that correlates small catchment observations of Q to climate variability. The learned relationships are then generalized to global scale. **iv)** The daily SWE observations from the GlobSnow v2 product (Takala et al., 2011; Luojus et al., 2014) represent snow variations in the Nothern Hemisphere, while the mostly snow-free Southern Hemisphere is not covered. Cell-timesteps with no snow are encoded as missing values. Thus, the product was gap-filled using 8 d snow cover fraction (SCF) from MODIS (Hall and Riggs, 2016), disaggregated to daily using nearest neighbor, to obtain a global coverage: A cell timestep in the SWE product was set to 0 if a) SCF $\pm$ 12 d was in average below 10 % and b) all SWE observations SCF $\pm$ 12 d were missing.

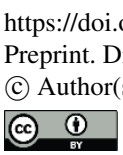
2.1.4 Data filtering

The grid cells were filtered to remove cases with 1) low variations in the hydrological cycle 2) high anthropogenic impact and 3) data limitations. 1) Grid cells with more than 50 % water bodies, more than 90 % permanent snow or ice, or more than 90 % bare land were removed. 2) Regions with high anthropogenic impact (e.g., groundwater withdrawal) as well as more than 90 % artificial surfaces were dropped. 3) Grid cells with more that 50 % missing values is the time series of the constraint variables were removed. The SWE product does not cover mountainous areas, which is also causing several grid cells to be removed. The filtered dataset contains a total number of 12 084 grid cells.

2.2 The hybrid hydrological model

The hybrid model (Fig. 1) consists three major blocks: a) the input data, b) the neural network module, and c) the hydrological model. The input data consists of the meteorological forcing time-series and the static variables (Sect. 2.1). The neural network yields a set of time-varying scalars which are used as model parameters in the hydrological model. The hydrological model represents major fluxes, such as snow accumulation and melt, soil recharge, groundwater recharge, runoff, and evapotranspiration, which are parameterized by the neural network.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021



Hydrology and Earth System Sciences
Discussions
Open Access
EGU

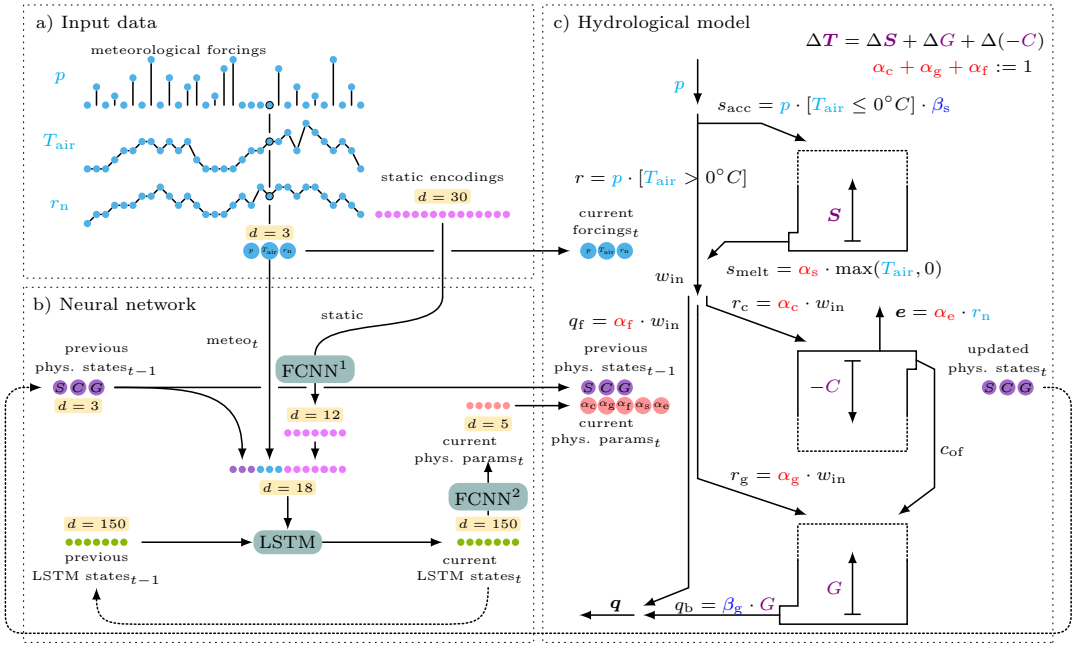


Figure 1. The hybrid hydrological model (H2M). The input data **a)** are fed into a neural network **b)** that estimates a set of scalars (parameters) used in a simple hydrological model **c)**. The input data **a)** consists of the meteorological variables precipitation (p), air temperature (T_{air}), and net radiation (r_{n}) and a set of encodings of static variables that are further compressed using a feed-forward neural network (FCNN^1). The neural network block **b)** contains a long short-term memory (LSTM) model at its core, receiving the input data from **a)**, together with the physical state variables snow water equivalent ($\boldsymbol{S}$), cumulative soil water deficit (C), and groundwater (G) at each time-step. The LSTM updates its hidden state at each time-step t (dotted arrows indicate recurrency). A second fully-connected neural network (FCNN^2) maps the LSTM state to the physical parameters soil recharge fraction (α_{c}), groundwater recharge fraction (α_{g}), fast runoff fraction (α_{f}), snowmelt coefficient (α_{s}), and evaporative fraction (α_{e}). These parameters are used as time-varying parameters in the hydrological block **c)**. The hydrological module updates the storage components $\boldsymbol{S}$, C, and G at each time-step: Snowfall (s_{acc}) is added to $\boldsymbol{S}$, while snowmelt (s_{melt}) is subtracted and added to rainfall, yielding the liquid water input (w_{in}). This quantity is partitioned according to α_{c}, α_{g}, and α_{f} into respective fluxes of soil recharge (r_{c}), groundwater recharge (r_{g}), and fast runoff (q_{f}). Note that s_{acc} is bias-corrected using a global parameter s_{corr}. Evapotranspiration ($\boldsymbol{e}$) is added to C (i.e., making the deficit larger), and as C approaches 0, exceeding water ('overflow', c_{of}) is passed to G. The baseflow (q_{b}) is simply groundwater G times a global constant β that is, together with the fast runoff, the total runoff ($\boldsymbol{q}$). The terrestrial water storage ($\boldsymbol{T}$) variations are computed as the sum of the variations in $\boldsymbol{S}$, G, and C. The updated states are passed forward to the LSTM (dotted arrows indicate recurrency). The boldfaced variables ($\boldsymbol{S}$, $\boldsymbol{T}$, $\boldsymbol{e}$, $\boldsymbol{q}$) are used to constrain the model with observations, upper case variables ($\boldsymbol{S}$, C, G) are states.

In this section, the model components is described in detail: The neural network is presented in Sect. 2.2.1, the hydrological balance equations in Sect. 2.2.2–2.2.6. In the hydrological balance equations, the time index t is implied and not noted explicitly

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

for fluxes and time-varying parameters. The symbol 'α' denotes time-varying scalars (parameters) directly estimated by the neural network, 'β' is used for constant, global parameters.

2.2.1 Neural network

The neural network (Fig 1b) consists of three modules. The first fully-connected model (FCNN[1]) has a single layer with 100 hidden nodes, followed by a leaky ReLU activation (Eq. 1). This layer reduces the static input variables θ from 30 to 12 values, yielding θ_{enc}. The long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) introduces the recurrency by maintaining two hidden states (h_t and c_t), each a vector of length 150, which are updated at every timestep (Eq. 2). Note that the cell state c_t is omitted in Fig. 1 for simplicity; c_t can be interpreted as the long-term memory, which is only used internally by the LSTM. In addition to the states, the LSTM receives a number of inputs: the physical states of snow S_{t-1}, cumulative soil water deficit C_{t-1}, and groundwater G_{t-1} from the previous timestep, the current meteorological forcings precipitation p_t, air temperature $T_{\text{air},t}$, and net radiation $r_{\text{n},t}$, together with the encoded static variables θ_{enc}, which do not change over time. The second fully-connected layer (FCNN[2]) with 50 hidden nodes maps the hidden state h_t to the physical parameters α_t, a vector of five scalars corresponding to α_c, α_g, α_f, α_s, and α_e, which are introduced later. For a comprehensive overview of the deep-learning methods used here, we suggest Goodfellow et al. (2016), available online at www.deeplearningbook.org.

$$\theta_{\text{enc}} = \text{ReLU}_{\text{leaky}}(\text{FCNN}^1(\theta)) \tag{1}$$

$$h_t, c_t = \text{LSTM}([h_{t-1}, c_{t-1}], [S_{t-1}, C_{t-1}, G_{t-1}, p_t, T_{\text{air},t}, r_{\text{n},t}, \theta_{\text{enc}}]) \tag{2}$$

$$\alpha_t = \text{FCNN}^2(h_t) \tag{3}$$

2.2.2 Snow

Snow accumulation s_{acc} (mm d^{-1}) is restricted to air temperatures T_{air} (°C) at and below the freezing point and corrected for the known overestimation of precipitation in the form of snowfall due to over-correction of snowfall undercatch by the factor $0 < \beta_{\text{s}} < 1$ (Eq. 4). Potential snowmelt s_{melt} (mm d^{-1}) is calculated using a degree-day approach and can only occur with positive air temperatures (Eq. 5). The time-varying snowmelt coefficient α_{s} is estimated by the neural network and mapped to the range $(0, \infty)$ by applying the sofplus function. The snow water equivalent S_t (mm) at time t is then updated using snow accumulation and melt, and negative values are prevented by truncating exceeding potential snowmelt (Eq. 6).

$$s_{\text{acc}} = p \cdot [T_{\text{air}} \leq 0] \cdot \beta_{\text{s}} \tag{4}$$

$$s_{\text{melt}} = \alpha_{\text{s}} \cdot \max(T_{\text{air}}, 0) \tag{5}$$

$$S_t = \max(S_{t-1} + s_{\text{acc}} - s_{\text{melt}}, 0) \tag{6}$$

2.2.3 Soil recharge, overflow, and evapotranspiration

The liquid phase water input w_{in} (mm d^{-1})—the sum of snowmelt and rainfall—, is partitioned into the three fluxes of surface runoff q_{f} (mm d^{-1}), soil recharge r_{c} (mm d^{-1}), and groundwater recharge r_{g} (mm d^{-1}). The parameters for the partitioning are

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions
Open Access
EGU

provided by the neural network and mapped to the range $(0, 1)$ as well as constrained to sum up to 1 by applying the softmax transformation. Soil recharge r_c ($\mathrm{mm\,d^{-1}}$) is, thus, a function of the liquid phase water input times the soil recharge partitioning α_c (Eq. 7). Evapotranspiration e ($\mathrm{mm\,d^{-1}}$) is calculated as the net radiation r_n ($\mathrm{MJ\,d^{-1}\,m^{-2}}$) converted to $\mathrm{mm\,d^{-1}}$ assuming a latent heat of vaporization of 2.45 ($\mathrm{MJ\,mm^{-1}\,m^{-2}}$), times the evaporative fraction α_e, which is learned by the neural network (Eq. 8) and mapped to the range $(0, 1)$ by applying the sigmoid activation. The soil moisture is parameterized as cumulative soil water deficit $C \geq 0$ (mm), which has the benefit of having a physical saturation limit of 0. The state C is updated by addition of the soil recharge, subtraction of e (Eq. 9), and leveling by the overflow mechanism (Eq. 10–11): If C approaches 0, an overflow mechanism redirects exceeding water input into the groundwater pool. Due to the heterogeneity within a model cell, the overflow c_{of} ($\mathrm{mm\,d^{-1}}$) starts already at values close to 0, which is achieved by using the $\mathrm{softplus}$ function.

$$r_c = \alpha_c \cdot w_{in} \tag{7}$$

$$e = \alpha_e \cdot \frac{r_n}{2.45} \tag{8}$$

$$C_t^* = C_{t-1} + r_c - e \tag{9}$$

$$c_{of} = \mathrm{softplus}(C_t^*) \tag{10}$$

$$C_t = C_t^* - c_{of} \tag{11}$$

2.2.4 Groundwater

Groundwater G (mm) is an unlimited storage that is refilled using two mechanisms (Eq. 14): The groundwater recharge, parameterized by the groundwater recharge fraction α_g (Eq. 12), and the overflow from the soil c_{of} (Eq. 10). Groundwater depletion happens via the baseflow q_b ($\mathrm{mm\,d^{-1}}$), which is the global constant β_g times the current groundwater state (Eq. 13).

$$r_g = \alpha_g \cdot w_{in} \tag{12}$$

$$q_b = G_{t-1} \cdot \beta_g \tag{13}$$

$$G_t = G_{t-1} + c_{of} + r_g - q_b \tag{14}$$

2.2.5 Runoff

The total runoff q ($\mathrm{mm\,d^{-1}}$) is parameterized as the sum of surface runoff q_f ($\mathrm{mm\,d^{-1}}$), and the baseflow q_b ($\mathrm{mm\,d^{-1}}$), shown in Eq. 16. Note that the neural network receives the storage states as inputs and is, thus, able to learn interactions of S_{t-1}, C_{t-1}, G_{t-1}, and input variables. Thus, the runoff generating processes can not only depend on the current meteorological forcing and the static variables, but also on, for example, the soil water deficit.

$$q_f = \alpha_g \cdot w_{in} \tag{15}$$

$$q = q_f + q_b \tag{16}$$

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

2.2.6 Constraint variables

The sum of the variation of the terrestrial water storage components yields the terrestrial water storage variations T (mm). Note that C denotes the water *deficit*, i.e., $-C$ is used in Eq. 17. Together with S, e, and q, ΔT is used for multi-objective model optimization.

$$\Delta T_t = \Delta S_t + \Delta G_t + \Delta(-C_t) \tag{17}$$

2.3 Model training

As the neural networks and the hydrological equations are differentiable, standard gradient descend approaches with backpropagation can be used for model optimization (Goodfellow et al., 2016). The model was implemented in *PyTorch* 1.5 (Paszke et al., 2017), an open source deep learning framework for the *Python* programming language. We followed a two-step procedure to 1) find a good set of hyper-parameters and 2) train the models in a cross-validation set-up. The global data was split into four different subsets (CV1–4) such that the spatial dependency between samples within a subset was reduced (although not completely removed). The grids were further randomly subdivided into five folds. In addition, the temporal domain was split into two periods, January 2002 to December 2008 for training and January 2009 to December 2014 for validation and test. For hyper-parameter tuning, we employed the Bayesian optimization hyper-band (BOHB) algorithm (Falkner et al., 2018) as implemented in the *ray.tune* framework (Liaw et al., 2018). The hyperparameter optimization and cross-validation procedure are described more detailed in Kraft et al. (2020).

To equilibrate the model's states (*i.e.*, S, G, C, and the LSTM hidden states), a spinup of five years was done using random years from the respective meteorological forcing data: in each optimization iteration, a forward model run on the spinup data was performed to retrieve steady states. These states were then used as initial states in the forward run on the actual training data, which included parameter updates.

The model was optimized on the four data constraint variables concurrently using the mean square error (MSE) as objective function. A further loss term was introduced to regularize the initial training phase: as already observed in previous experiments (Kraft et al., 2020), the mean C was not properly constrained. We hypothesize that the state drift originates from the spinup procedure, where the randomly initialized neural network parameters lead to erratic behavior in the early training phase. To reduce the state drift, a loss term was introduced to push the lower 10 percentile of C towards 0. This loss term was gradually given less weight during training. The five loss terms were dynamically weighted using self-paced task weighting approach proposed by Kendall et al. (2018)—we refer to Kraft et al. (2020) for more details.

2.4 Model evaluation

2.4.1 Performance metrics

The quality of the model predictions was assessed using different metrics. The Nash–Sutcliffe model efficiency coefficient (NSE, Eq. 18–19), was adapted by transforming negative values from the range $(-\infty, 0)$ to a range of (-1, 0) in order to avoid

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

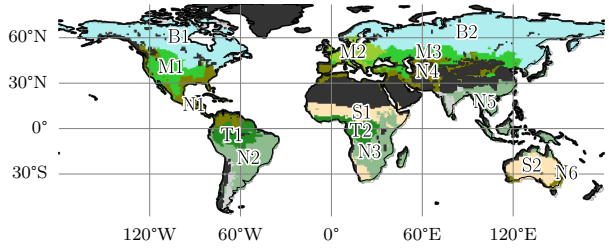


Figure 2. Continental hydro-climatic regions, adapted from Papagiannopoulou et al. (2018). **Boreal**: North America (B1) and Eurasia (B2). **Temperate**: North America (M1), Europe (M2), and Asia (M3). **Transitional**: North and Central America (N1), South America (S2), Africa (N3), Eurasia and North Africa (N4), Southeast Asia (N5), and Australia (N6). **Subtropical**: Africa (S1) and Australia (S2). **Tropical**: South America (T1) and Africa (T2).

large negative values (Nash and Sutcliffe, 1970).

$$\mathrm{NSE}^* = 1 - \frac{\sum_{i=1}^{N} (m_i - o_i)^2}{\sum_{i=1}^{N} (o_i - \bar{o}_i)^2} \tag{18}$$

$$\mathrm{NSE} = \begin{cases} \mathrm{NSE}^*, & \text{if } \mathrm{NSE}^* \geq 0 \\ \tanh(\mathrm{NSE}^*), & \text{otherwise} \end{cases} \tag{19}$$

where m_i is the modeled and o_i the observed value at sample i of N samples, $\bar{o}$ is the mean of the observed time-series. Further, the root mean square error (RMSE), the Pearson correlation (r), and the ratio of standard deviation (SRD) were used.

2.4.2 Temporal and spatial scales

The observed and simulated time-series were decomposed independently into the mean seasonal cycle (MSC) and the interannual variability (IAV):

$$V_{\mathrm{MSC}_m} = \frac{1}{Y} \sum_{y=1}^{Y} V_{m,y} \tag{20}$$

$$V_{\mathrm{IAV}_{m,y}} = V_{m,y} - V_{\mathrm{MSC}_m} \tag{21}$$

where V is the observed or modeled time-series, m is the month index and y is a year out of Y total years. For calculating the model performance, the linear trend was removed before computing the MSC and IAV, but not from the raw time-series V. This was done because the calculation of the MSC and the IAV can be strongly affected by linear trends.

We evaluate the model performance on different spatial scales to emphasize both the local variations and the coarser scale dynamics. For this, we split the spatial domain continent-wise into similar hydro-climatic regimes (Fig. 2). To account for the varying cell areas, all reported aggregated metrics were weighted by the cell area.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

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 (if available) and groundwater plus surface storage (if available), respectively.

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

SWE=soil water equivalent, CWD=cumulative water deficit, GW=groundwater, SM=soil moisture, CInt=canopy iterseption, GW=groundwater, SStor=surface storage

2.5 Global hydrological model ensemble

To evaluate the simulations of TWS and its components (SWE, CWD, and GW), we contrast them to a selection of models from the eartH2Observe model ensemble (Schellekens et al., 2017). From the ten available models, we select those for which a groundwater estimate is available: 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). The four models represent different components of the terrestrial water storage (Table 2). To compare the model storage components, we calculated CWD as the soil moisture (SM) plus canopy interception (if available) and groundwater plus surface storage (if available). This approximation has certain implications, which are to be considered in the discussion of the results. We consider the dynamics of CWD to correspond to SM and thus, the terms are used interchangeably when talking about soil moisture dynamics.

2.6 Terrestrial water storage decomposition

We use the model simulations of the variables CWD, GW, and SWE to assess their contributions to the TWS dynamics, seasonality, and interannual variability. Note that the model does not represent surface water storage—a fourth component of TWS—explicitly. This will be considered in the discussion of the results. The absolute contribution $\mathcal{A}$ is calculated, follow-

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

ing Getirana et al. (2017), as:

$$\mathcal{A}_{v\in\{C,G,S\}} = \sum_{t=1}^{\tau} |V_{v,t} - \bar{V}_v| \tag{22}$$

$$\mathcal{C}_{v\in\{C,G,S\}} = \frac{\mathcal{A}_v}{\sum_{w\in\{C,G,S\}} \mathcal{A}_w} \tag{23}$$

where $\bar{V}_v$ is the mean over the time-series V_v, and $\mathcal{C}_v$ is the relative contribution (hereinafter simply *contribution*) of a component v. The contributions are calculated grid cell wise for the mean-removed monthly time-series V_t, and their decomposition into MSC and IAV. Note that negative C was used in Eq. 22 as the values indicate a *deficit* in soil water. Throughout the manuscript, we use cyan to represent SWE, yellow for CWD and soil moisture dynamics, and magenta for groundwater. We tried to use colorblind friendly colors in the illustrations whenever possible.

3 Results

We first assess the hybrid model performance on different spatial and temporal scales in respect to the four data constrain variables (TWS, SWE, Q, and ET), followed by a comparison to the four process models, where the common variables TWS and SWE are evaluated. As several changes were made to the model since Kraft et al. (2020), we will re-evaluate its performance here, in more detail. We then take a closer look at the parameters estimated by the neural network that define the hydrologic responses and generation of key hydrological fluxes. Finally, we investigate how the different models partition TWS into the components of snow, soil moisture and groundwater.

3.1 General model performance

The model reproduced the patterns well (Tab. 3). In general, the spatially averaged signal was reproduced better than the cell median. For both observational constraint variables TWS and SWE an NSE>0.8 and r>0.9 for the averaged signal and a cell median NSE>0.5 and r>0.8 was achieved. The seasonal signal TWS_{MSC} and SWE_{MSC} were modeled with high accuracy (NSE>0.9 for averaged, NSE=0.7 for median cell level) while the interannual variability performance varied: The TWS_{IAV} was reproduced well with NSE=0.54 (r=0.8) for the spatial average, and a median cell NSE=0.26 (r=0.67). The SWE_{IAV} performance was decent for the averaged signal (NSE=0.22, r=0.87), but lower (NSE=0.15, r=0.64) on median cell level.

Both ET and Q, which are machine learning model based and not directly observed at global scale, were reproduced well in terms of the seasonality on the spatial averaged signal, while the cell level performance was lower. For the ET_{IAV}, low NSE=-0.17 on global, and NSE=-0.65 on cell level is achieved, while the correlation is still relatively good with r=0.67 on global, and r=0.6 on cell level. The SDE indicates that on both global and cell level, the variability of the simulated ET_{IAV} signal is substantially larger than the reference data with SDE=1.41 on global, and SDE=1.65 on cell level (see Fig. A2 in the Appendix for spatial patterns). For Q, the performance is decent on global level and lower on cell level. Also here, low values in terms of NSE are accompanied with relatively good correlation. Because the independent data for ET and Q are not direct observations,

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions

Table 3. Model performance of the monthly spatially averaged and the median cell-level performance for the observational constraint variables terrestrial water storage (TWS) and snow water equivalent (SWE), evapotranspiration (ET), and runoff (Q), their decomposition into the mean seasonal cycle (MSC) and interannual variability (IAV). The metrics 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, values represent the mean across the 15 cross-validation runs. Positive values of SRD indicate that the modeled variance is larger than the observed. Note that for the SWE, cells with constant 0 were dropped.

		TWS			SWE			ET			Q		
	Metric		MSC	IAV		MSC	IAV		MSC	IAV		MSC	IAV
Spatial mean	NSE	0.84	0.93	0.54	0.96	0.96	0.22	0.96	0.96	-0.17	0.75	0.78	0.47
	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
Cell-level median	NSE	0.54	0.70	0.26	0.58	0.74	0.15	0.79	0.87	-0.65	0.20	0.17	0.07
	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

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 procided in Appendix A.

3.1.1 Model intercomparison

We compare the simulations of the H2M to a set of state-of-the-art GHMs. We note here upfront that H2M was optimized with the datasets we analyze, while the GHMs have either been calibrated using catchment-level observational runoff data (LISFLOOD) or rely on prior parameter estimation (W3RA, SURFEX-TRIP, RCR-GLOBWB) alone (Schellekens et al., 2017). The question of the comparison is not "which model is better overall" but which features are relatively better or worse modeled across models. Note that the H2M model performance may differ from the numbers presented in the previous section, as the time-period from 2003 to 2012 was used for the model comparison because of model data availability.

The H2M modeling efficiency is higher than the GHMs' on local cell-level, while it falls within the range of the GHMs on spatially averaged scale. Figure 3 shows the global performance of the H2M and the GHMs contrasted. In terms of the spatially averaged TWS signal ($\diamond$ in Fig. 3), H2M and PCR-GLOBWB perform better than the other GHMs. While the PCR-GLOBWB reproduces the seasonality slightly better than the H2M, the latter performs better when it comes to the IAV. On the local scale (boxes in Fig. 3), the H2M outperforms the GHMs on the TWS, TWS_{MSC} and TWS_{IAV}, when comparing the median across cells. The SWE_{MSC} is reproduced best by H2M on both spatially averaged and local scale. All models struggle to reproduce

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

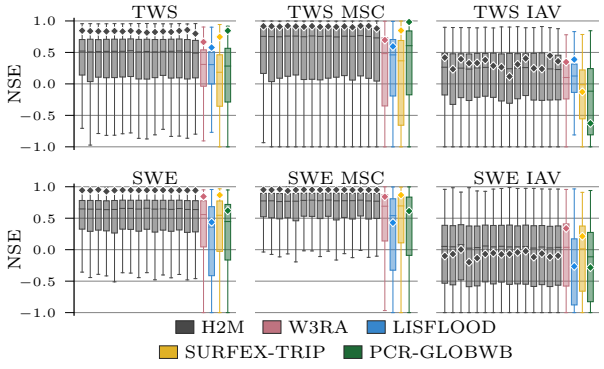


Figure 3. Globally averaged and local grid cell level Nash–Sutcliffe model efficiency coefficient (NSE) of the hybrid hydrological model (H2M) and the process-based global hydrological models (GHMs) for the terrestrial water storage (TWS) on top and the snow water equivalent (SWE) on bottom. The grey bars represent the cross-validation runs. The ⋄-markers show the global, spatially averaged model performance, the boxes represent the spatial variability of the cell level performance. The NSE is transformed to the range -1 to 1 by applying the hyperbolic tangent function to negative values due to large negative values occurring in some models. 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 one day 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 2003 to 2012. Note that deviations from the numbers reported in Tab. 3 are due to different time ranges.

the SWE_{IAV} signal: The median NSE of H2M is on a par with W3RA and SURFEX-TRIP, while the performance on spatially aggregated level is lower.

Figure 4 shows the zonal distributions of phase and variance error of H2M compared to the GHMs. The H2M performance is usually within the range and often at the lower end of the GHM errors. The zonal distributions of errors from GHMs and H2M are similar, which suggests that both have lower performance in the same geographical regions where the data uncertainties may be large or process representations and not suitable and sufficient. The largest variance errors for TWS occur in the tropical and subtropical zones and in very high latitudes. The high latitude variance error is also present in the SWE.

On regional scale, most models reproduced the TWS_{MSC} well (NSE>0.5), while the TWS_{IAV} performance varied (NSE<0.5) (Fig. 6). 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), 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_{MSE} well with an NES>0.5, while the SWE_{IAV} performance is more variant and lower in general (Fig. 6). 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

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

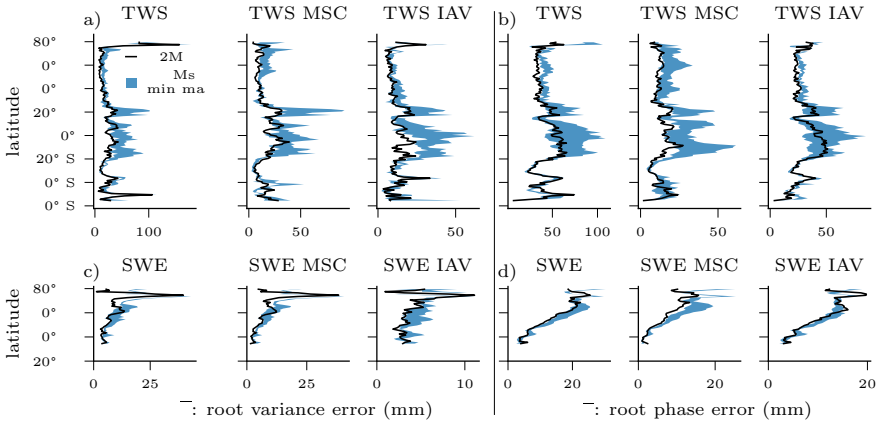


Figure 4. Comparison of the root phase and variance error of the hybrid hydrological model (H2M) to the process-based global hydrological models (GHMs) for terrestrial water storage (TWS) and the snow water equivalent (SWE) time-series, their seasonality (MSC) and interannual variability (IAV). Plot a) shows the TWS variance error, b) the TWS phase error, c) the SWE variance error, and d) the SWE phase error. The black line represents the mean latitudinal error (average across longitudes) of the H2M and the shaded area is the minimum to maximum error of the GHMs. The metrics are calculated for the test period from 2003 to 2012. Note that x-scale differs between plots.

315 across different regions. The SWE_{IAV} has strong seasonal variations, with opposite patterns in different regions that cancel each other out on 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.

3.2 Hydrological responses

The H2M yields a set of data-driven, spatio-temporally varying estimates of model parameters that define the hydrologic re-
320 sponses and generation of key hydrological fluxes. In particular, we focus on four parameters, α_c, the fraction of throughfall that percolates into soil, α_g, the fraction that recharges the groundwater, α_f, the fraction that runs off as surface runoff component, and α_e, the evaporative fraction (ratio of evapotranspiration to net radiation). In this section, we analyze the spatiotemporal variability of these parameters and how they are associated with the antecedent moisture condition defined by soil water deficit (larger CWD).

325 The partitioning of the liquid water input w_{inp} (rainfall plus snowmelt) using the fractions for soil recharge (α_c), groundwater recharge (α_g), and surface runoff (α_f) was robust across cross-validation runs and showed a clear relationship to CWD (Fig. 7). With an increasing soil water deficit (larger CWD), the soil recharge increases, while the groundwater recharge and surface runoff decrease. For a CWD<200 mm, we see a large spatio-temporal variation in the partitioning, evident through the relatively large difference between the 0.2 and the 0.8 quantile. The transition from larger soil recharge to larger groundwater recharge

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

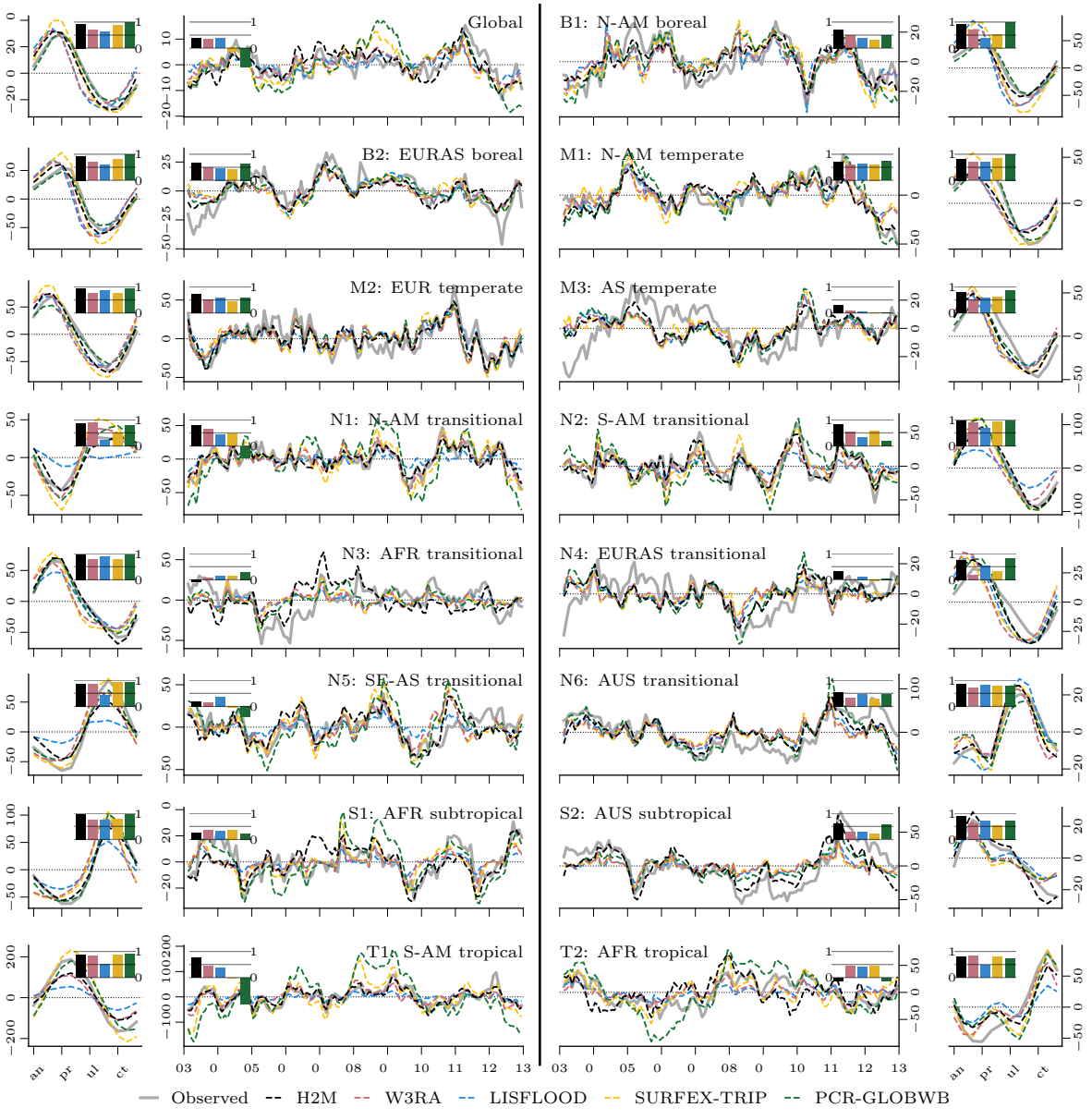


Figure 5. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the terrestrial water storage mean seasonal cycle (TWS_{MSC}) and interannual variability (TWS_{IAV}) in mm for hydro-climatic regions (Fig. 2). The time-series were aggregated 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.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

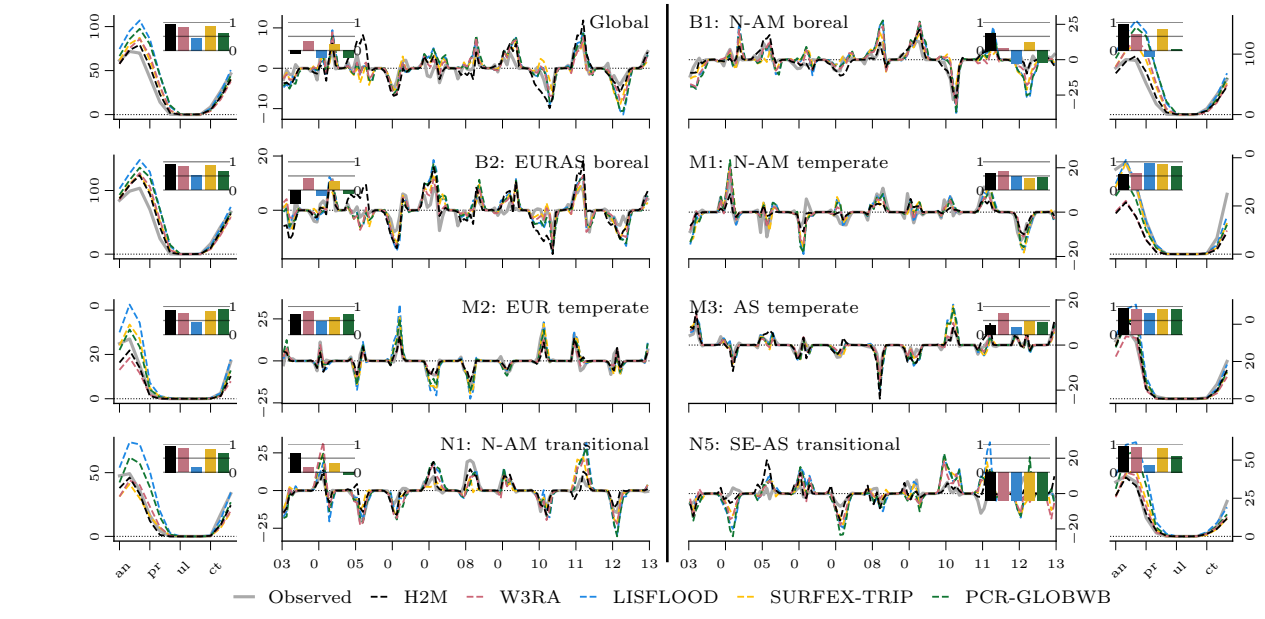


Figure 6. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the snow water equivalent mean seasonal cycle (SWE_{MSC}) and interannual variability (SWE_{IAV}) in mm for hydro-climatic regions (Fig. 2). The time-series were aggregated using 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. Note that the y-scale differs between plots.

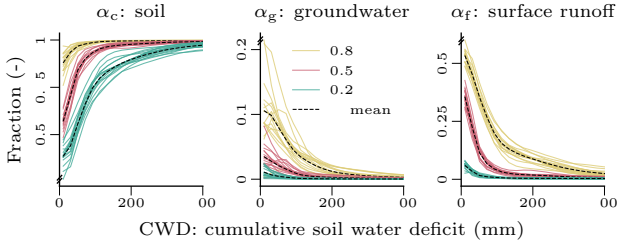


Figure 7. Relationship between the water input partitioning fractions for soil (α_c), groundwater (α_g), and fast runoff (α_f) and the cumulative soil water deficit (CWD) as learned by the neural network. The colored lines represent the 0.2, 0.5, and 0.8 quantiles of the spatio-temporal distribution for different cross-validation runs to show the robustness of the simulations. The dashed, dark lines are the average across the runs per quantile. The plots are based on global daily cell-timesteps from 2009 to 2014. Note that the y-scale differs between plots.

and surface runoff is exponentially decreasing, i.e., the change is faster with lower CWD. Above a CWD of 200 mm, the partitioning is constant in space and time with α_c converging to 1, while α_g and α_f converge to 0.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

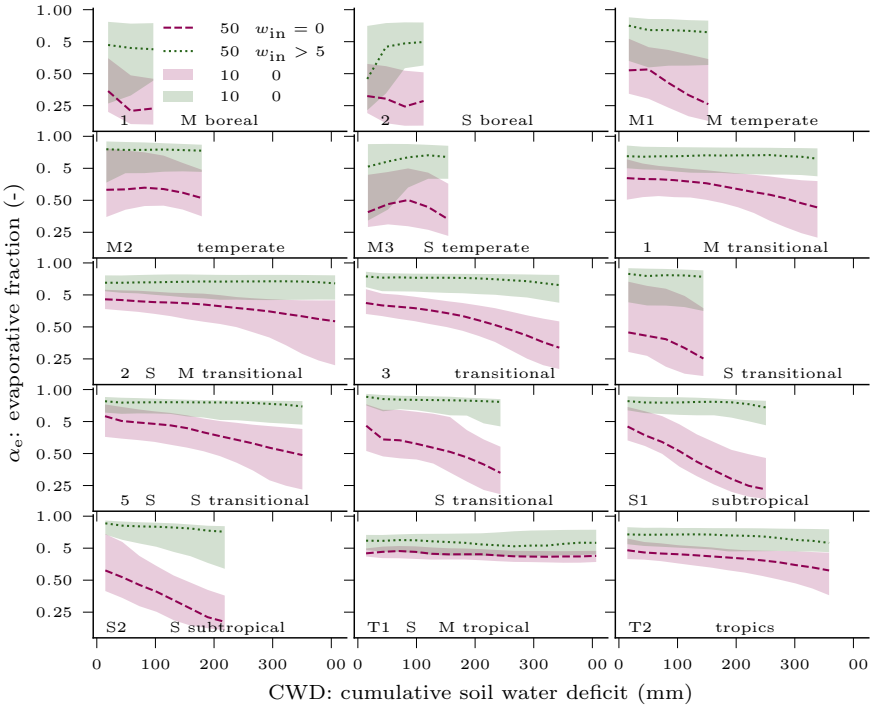


Figure 8. Relationship between evaporative fraction (α_e) and cumulative soil water deficit (CWD) for different hydroclimatic regions. The median and 0.1–0.9 quantile range is shown for conditions without water input ($w_{\mathrm{in}} = 0$ mm), i.e., no precipitation or snowmelt, and with high water input ($w_{\mathrm{in}} > 5$ mm). 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 percent maximum CWD were used

In most hydroclimatic regions, the α_{e} showed a negative relationship to CWD under dry conditions, and no relationship in presence of precipitation or snowmelt (Fig. 8). The high latitude and tropical regions showed a less clear relationship and less variation in CWD in general. In all regions, α_{e} was close to 1 with large water input ($w_{\mathrm{in}} > 5$ mm). In arid and semiarid climates, α_{e} takes a larger range of values, decreasing with CWD. The 0.1–0.9 quantile spread is large in most cases, which indicates that the relationship is modeled with a large spatio-temporal variability.

The mean α_{e} shows hotspots in temperate and tropical regions, while lowest values are in arid and semiarid climates (Fig. 9). The Figure also shows the relationship between α_{e} and CWD in more detail for certain locations. The boreal site in Northern America (A) shows low α_{e} around 0.2 on average and no interaction with soil moisture, and a similar relationship yet with a generally larger α_{e} is found in Eastern Europe (C), with values in the range 0.4 to 0.9. In the Amazon basin (B), we do not see an interaction between α_{e} and CWD as well, the values are generally large (0.8–0.9), and precipitation has only a small impact on the relationship. These regions are characterized by low soil moisture variations with a maximum CWD of 100 to 200 mm. For the remaining sites, South Africa (D), India (E), and East Australia (F), a clear relationship between soil water stress and α_{e} is found. While under wet conditions α_e is close to 1, dry conditions (low precipitation) lead to a decrease in α_{e} from around

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

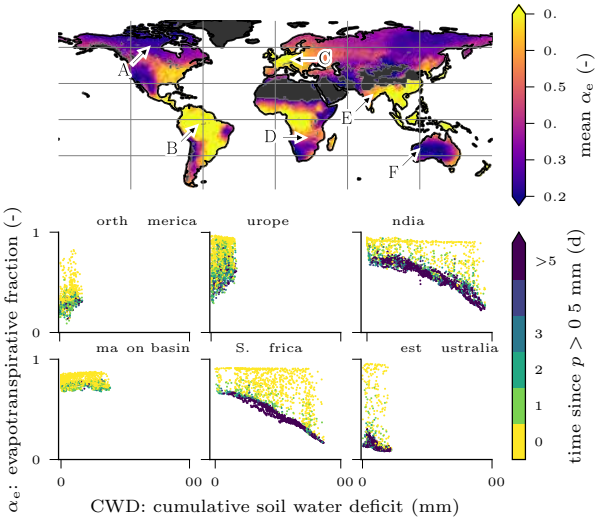


Figure 9. The map shows the mean evaporative fraction (α_e) and scatterplots display the relationship between (α_e) and the cumulative water deficit (CWD), colored by days since last precipitation ($p > 0.5$ mm). The plots are based on global daily cell-timesteps, filtered for positive air temperatures and net radiation, from 2009 to 2014.

0.7 with saturated soil to 0.3 with dry soils. The E) India and D) South Africa regions show the largest CWD variations with a maximum close to 600 mm. All locations show a strong increase in α_e with significant precipitation ($p > 0.5$) on the same day, and less expressed, during the previous days.

3.3 Terrestrial water storage decomposition

In this section, we show the TWS partitioning into snow, soil moisture, and groundwater variations as done by the H2M model and compare the patterns to the ones of the GHMs.

The spatial patterns of the TWS partitioning vary strongly among the models (Fig. 10, top). Some patterns are consistent, though: The TWS seasonality is dominated by the SWE signal 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. Otherwise, the models show large discrepancies. Both W3RA and PCR-GLOBWB show 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 Rainforest in Amazon and Southeastern Asia show a distinct groundwater signal. For the TWS_{IAV} decomposition, we see a rough agreement between the H2M, LISFLOOD, W3RA, and PC-GLOBWB model in North America, Europa and northern and central Asia, while the latter two again show a stronger groundwater contribution, which extends to southern tropical and mild climates (Fig. 10, bottom). The strongest difference between H2M and the GHMs is the low groundwater IAV in Africa, which we also observed in the TWS_{MSC} decomposition.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

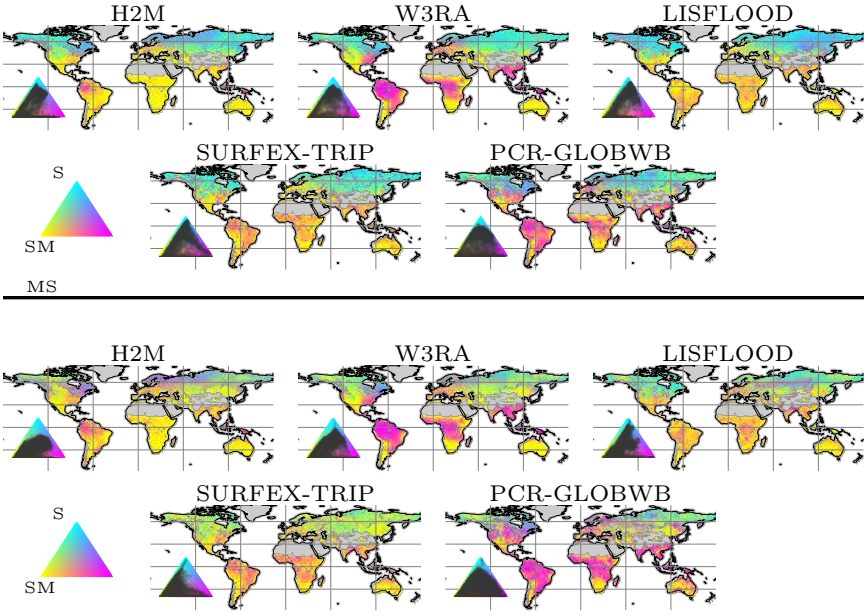


Figure 10. Terrestrial water storage (TWS) variation partitioning into cumulative water deficit (CWD), groundwater (GW), and snow water equivalent (SWE) variation based on the validation period for the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs). The top panels show the partitioning of the mean seasonal cycle (MSC), the bottom the interannual variability. The map colors correspond to the mixture of the contributions of the two variables, 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.

Not only the spatial patterns of the TWS partitioning shows large variations. The global signals of the components, shown in Fig. 11, illustrates the differences in amplitude and timing for the time-series and their decomposition into MSC and IAV. For the seasonal TWS signal, the amplitudes are qualitatively similar, and the main contribution comes from the snow pack. H2M, SURFEX-TRIP, and PCR-GLOBWB show a soil moisture slightly delayed to the snow seasonality, and the groundwater peak setting in in the late northern spring. W3RA shows very similar soil moisture and groundwater curves, being 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 similar, but the amplitudes largely differ across the modes. 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.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

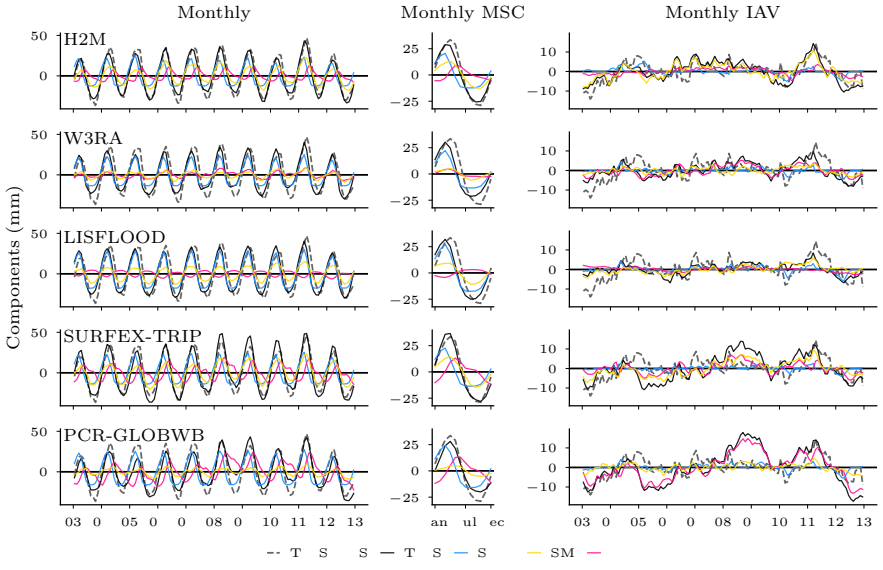


Figure 11. Global average variability of the terrestrial water storage (TWS) and the components snow water equivalent (SWE), soil moisture (SM), and groundwater (GW) for the hybrid hydrological model (H2M) and the process-based global hydrological models (rows). For reference, the TWS observations are shown (TWS OBS). The monthly signal (left) and its decomposition into the mean seasonal cycle (MSC, center) and the interannual variability (IAV, right) are shown (columns). The time-series were aggregated using the cell size weighted average, only cell-timesteps 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.

The regional scale seasonal anomalies of simulated SM and GW show a more detailed picture of the model variabilities (Fig. 12). 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 North (N1) and South (N2) America transitional, the Australia subtropical (S2), and Africa 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) as well as in the Africa (N3), Eurasia (N4), and Australia (N6) transitional zones.

4 Discussion

In this section, we discuss the plausibility and implications of a set of hydrological responses simulated by H2M. First, the learned relationship between CWD and runoff generating processes is discussed, followed by an analysis of the CWD-α_e (evaporative fraction) relationship. Then, the TWS composition by H2M is contrasted to GHM simulations.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

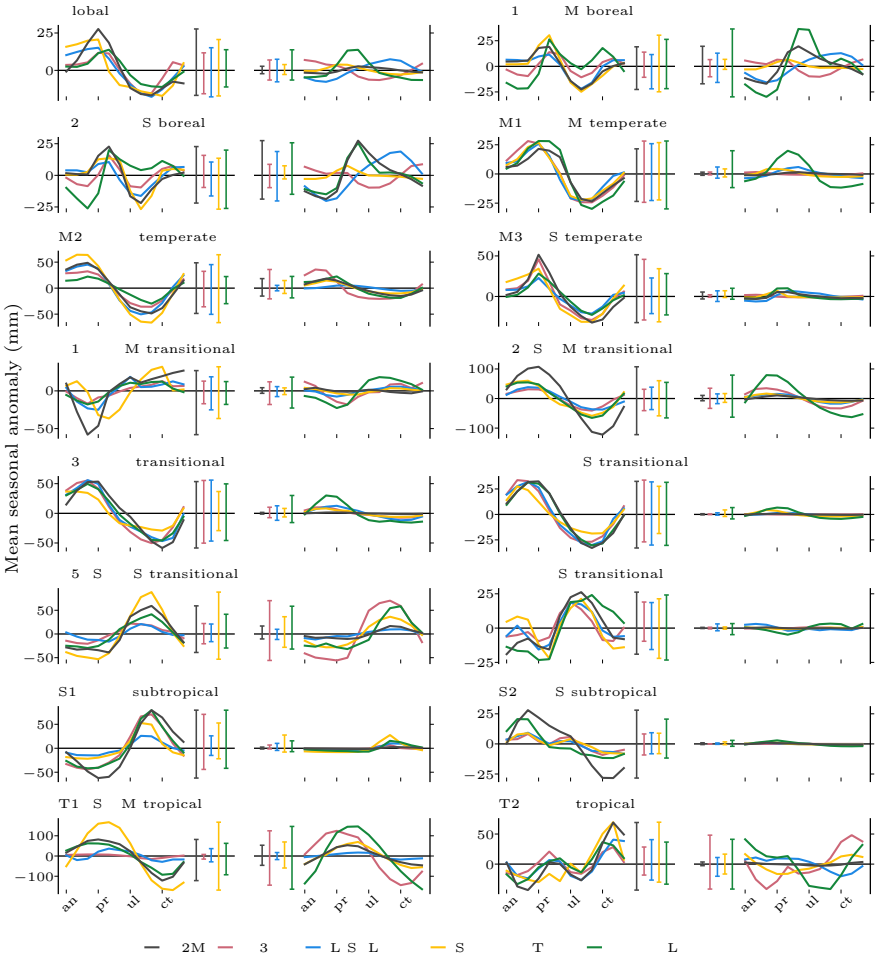


Figure 12. Global and regional mean seasonal anomalies of soil moisture (SM) and groundwater (GW) for the hybrid model (H2M) and the process-based global hydrological models. Ranges from the minimum to the maximum value per model are shown next to the seasonal cycle as vertical lines. The regions are shown in Figure 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-timesteps from 2009 to 2014. Note that the y-scale is consistent within, but differs across regions.

4.1 Model performance

The H2M model simulations have a good agreement with the TWS and SWE observations given the data biases and the rather simple hydrological balance equations that were used to constrain the recurrent neural network. The TWS seasonality was reproduced well, 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 r. Largest RMSE are in humid regions with stark TWS seasonality and large runoff

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

rates, e.g, the Amazon basin, central Africa, and Southeast Asia (Fig. A1). This may be related to the missing representation of delayed water storage, e.g., due to lateral flow, that are dominant in humid river basins (Kim et al., 2009). As a consequence of the missing fluvial transport, the cells cannot receive water input from their neighbors. At the same time, the H2M model does not implement surface water storage and thus, there is no explicit mechanism to represent buffering as it happens in wetlands. Thus, the seasonal amplitude of TWS is underestimated by the H2M in wetlands with strong lateral fluxes (also see Fig. A2). The models with surface water storage representation (SURFEX-TRIP and PCR-GLOBWB) manage to better reproduce the seasonal amplitude in the Amazon, even if this is not the case for Southeast Asia and central Africa (Fig. 5). The importance of representing surface water storage has been highlighted before (Scanlon et al., 2019) and should be considered in further development of the H2M model. A further source of errors are signals of human intervention, such as irrigation, that cannot be picked up by the model.

For the SWE seasonality, a near-perfect fit was achieved for the globally averaged signal (Fig. 5). Locally, however, the performance varied strongly across regions with the poorest performance in extremely cold tundra. This is possibly linked to two factors. First, the globsnow SWE observation saturates at around 100 to 120 mm, which has been shown for both North America (Larue et al., 2017) and Eurasia (Luojus et al., 2010). Consequently, the H2M overestimates the mean SWE in both regions (see also Fig. A1 and A2). Second, the GPCP precipitation overestimates snowfall due to over-correction of snowfall undercatch, especially in regions with low density of in-situ measurements such as the high latitudes with large local biases (Behrangi et al., 2016; Panahi and Behrangi, 2019). To account for this, we introduced a snowfall correction factor, estimated as $\beta_\mathrm{s} = 0.77 \pm 0.01$ (cross-validation mean and standard deviation). This global correction factor may reduce biases over all regions, but does not address the differences in regional biases. For the SWE interannual variability, we see similar, yet larger regions of lowered model performance. Due to the saturation of the globsnow product, the true interannual variability is likely to be underestimated in the observational product, especially under the presence of a large SWE.

The H2M performance aligned well with the process-based GHMs. On the spatially averaged and regional level, H2M performs on par with the best GHM, while the overall grid cell-level performance is even better than GHMs. This highlights the key strength of the hybrid approach: the local adaptivity. Only for the SWE interannual variability, the performance of H2M is not better than the GHMs on the grid cell level. The hybrid model represents the snow processes in a relatively rigid way, allowing snowfall only below 0°C, and snowmelt above. This reduces the data adaptivity largely, as preliminary experiments have shown, but increases the physical consistency of the model. By increasing the physical accuracy, the model loses its flexibility to compensate for data biases, similar to the GHMs. Similarly, the other hydrological constraints limit the flexibility of the model. While this is needed to obtain interpretable estimates of parameters and hydrological responses, wrong or simplistic process representations lead to a lowered performance but also to compensation effects in the model, and ultimately inflect the interpretable variables. Thus, further development of the H2M model should focus on a more accurate representation of the hydrological processes to reduce model biases.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

4.2 Model interpretability

In this section, we assess the model interpretability, i.e., the plausibility of the hydrological responses and parameters. First, the partitioning of precipitation and snowmelt into soil moisture recharge, groundwater recharge, and surface runoff is discussed. Next, we look at interactions of CWD and evaporative fraction, and finally, the plausibility of the TWS partitioning is evaluated.

4.2.1 Hydrological responses

The H2M model learned hydrological responses to soil moisture status that are consistent with our understanding. The partitioning of incoming water in surface runoff and recharge of the soil and groundwater shows a clear non-linear response to CWD (Fig. 7). The fraction of surface runoff (α_{f}) decreases rapidly with increasing dryness while soil recharge (α_{c}) increases correspondingly. This runoff generating process response to soil moisture matches qualitatively the expected behavior implemented in GHMs (Bergström, 1995).

The H2M predicts large spatial-temporal variability of the soil moisture dependent runoff-recharge partitioning as indicated by different quantiles in Fig. 7. For example, under moist conditions (low CWD) more than 50 % of water input (blue lines in Fig. 7) or hardly anything (yellow lines) can be directed to fast runoff. Also the CWD point at which the runoff-recharge fractions level off appears to vary substantially. 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. Representing this large natural variability in a process-oriented manner has been a key challenge in traditional GHMs primarily due to uncertainties of representing the effective behavior of sub-grid variability associated with heterogeneous landscapes, complex processes and dynamics (Döll and Flörke, 2005; Beck et al., 2016, 2017; Koirala et al., 2017). Therefore, parameters of this hydrological response are typically "effective" calibration parameters in GHMs, i.e., parameters that describe the mean behavior that do not have a direct physical interpretation. Here, the H2M approach offers interesting perspectives in modeling and better understanding the effective runoff generating process response to soil moisture by learning the effective behavior constrained by multiple observation data streams.

Groundwater recharge in H2M happens via two processes: 1) a simple bucket overflow dynamics where all water that cannot be retained in the soil (CWD close to 0) drains to the groundwater pool, and 2) a fraction of incoming water is directed to groundwater recharge (α_{g}) when the soil moisture is below field capacity (CWD>0). The latter process may capture groundwater recharge through preferential macro pore flow paths acknowledging the heterogeneity of soil hydraulic conductivity. In addition, spatial sub-grid variability of runoff-runon dynamics and moisture convergences can lead to groundwater recharge that cannot be simply represented by a vertical discretization of the soil alone. Interestingly, the response learned by H2M suggests that groundwater recharge at soil moisture below field capacity seems not very relevant overall (Fig. 8). The median groundwater recharge fraction is only a few percent at CWD=0 and converges to zero with increasing soil dryness. This suggests a correspondingly small role of complex sub-grid processes in generating groundwater recharge at coarser scales. Yet, the model structure and observational constraints of H2M may be insufficient here to state a robust claim though and further investigation is needed.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

The learned relationship between evaporative fraction (α_e) and soil dryness (Fig. 8 & 9) 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 α_e-CWD response functions appear to be rather gradual as opposed to an idealized piecewise function with a clear soil moisture threshold that is also frequently employed in process-models (Seneviratne et al., 2010; Schwingshackl et al., 2017). The specific response predicted by H2M varies substantially between regions and within regions indicated by the shading in Fig. 8. For example, α_e starts declining already at low dryness in semi-arid regions of Africa and Australia while α_e remains high at large moisture deficits in tropical regions. The large sensitivity of α_e to soil moisture in semi-arid regions is consistent with large fractions of herbaceous vegetation with shorter rooting depth there that respond very dynamically to moisture variations (Sperry and Hacke, 2002; Fan et al., 2017). In contrast, the α_e-CWD response in the wet tropics appears to be absent in South America and weak in Africa, which could be related to large storage capacities due to deep rooting of tropical forests such that soil dryness has not reached levels with associated water stress. In addition, shallow water tables that are widespread in the wet tropics (Fan et al., 2013) may support ET and alleviate water stress. For such conditions the conceptualization of CWD as a soil moisture pool from which ET is taken up in H2M would be misleading since plant water uptake from groundwater and capillary rise are not represented explicitly.

Vegetation storage capacity has long been identified as a key uncertainty in process-models in controlling soil moisture stress responses (Ichii et al., 2009). But in addition many factors are contributing to the large variability of the ET soil moisture stress response in nature. Soil properties control resistances and matrix potential in interaction with root and plant hydraulic traits while functional biodiversity was shown to be important as well (Sperry and Hacke, 2002; Fischer et al., 2019). Thus, the large uncertainty in representing this response in coarse process-models makes the machine learning approach in H2M very attractive, which is supported by better performance of H2M in simulating TWS variations in tropical and subtropical regions compared to GHMs (Sect. 3.1) despite its simple overall structure.

H2M predicted another intriguing feature of the evaporative fraction: an about constant potential EF was predicted when there was substantial rain, independent of the soil moisture state (green lines in Fig. 9 and dark points in Fig. 10). Thus the model implicitly accounts for wetting of the top soil layers which alleviates water stress even though it represents soil moisture as a single bucket. Such response cannot be represented in process-models without vertical discretization of the soil and suggests an effective and computationally cheap way of dealing with such processes by suitable machine learning approaches.

4.2.2 Terrestrial water storage composition

As reported previously (Andrew et al., 2017) and as presented here, the attribution of TWS variations is an outstanding challenge in global hydrological modeling. The cross-comparison of the data-driven hybrid approach against the spatio-temporal patterns from GHMs provides complementary insights into TWS variability.

The dominant contribution of the SWE in the high latitudes to the seasonal cycle of TWS (Fig. 11 & 12), but a lower contribution to the interannual variability is consistent across models, and also has 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

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

uncertainties in the 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. 10). 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 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 the interannual global scales, groundwater contributions to TWS correlate with humidity (c.f., 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. 11 & 12) co-occurring with low evaporative demand due to low temperatures, irradiation, and vegetation productivity, that 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 top soil layers 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 scale 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 Asia rainforests are the most humid ones globally with largest annual precipitation (Zelazowski et al., 2011) and a shallow plant rooting depth, while the African rainforest is somewhat dryer 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.

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 are a unique feature of H2M.

GHMs require a large number of parameters that are either empirically derived or based on remote sensing or statistical datasets, for example, plant functional types, root zone depth, or soil properties. Often, the said parameters are uncertain and may not 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 deeper layer variability 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;

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Koirala et al., 2017). Compared to the GHMs, H2M provides a novel avenue on which storage variations are not bound by the ad-hoc prescription of the size of soil and other storages. The diagnosed patterns of soil and groundwater variations therefore emerge from observation-based variations of water storage and fluxes. The H2M approach that also implicitly learns 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, that 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 represents 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 a larger soil moisture in transitional and especially in the subtropical regions, but more evidently, smaller groundwater variability. This effect may be reinforced by biases in the observational constraints, like an overestimation of ET by the remote sensing based FLUXCOM product (Tramontana et al., 2016; Jung et al., 2019) and large uncertainties of the precipitation data due to limitations in density and quality of measurement sites (Sylla et al., 2013) in Africa. These biases to can lead to smaller availability of moisture for recharge to groundwater storage, and lead to smaller variability of groundwater storage

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. Although H2M may implicitly represent delays associated with surface storage variation by assigning it to other storage components, the current implementation does not allow to diagnose and validate that explicitly. 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 Uncertainties, challenges, and opportunities

In this section, we discuss uncertainties emerging from the data constraints and the modeling approach, as well as outstanding challenges and opportunities.

4.3.1 Uncertainties

The hybrid modeling approach heavily depends on the quantity and quality of data used for forcing, characterizing land surface, and for constraining the model. Uncertainties in the forcing datasets have been found to strongly affect physically-based models, with precipitation having the largest uncertainties and impact on model quality (Döll et al., 2003; Beck et al., 2016). The hybrid model is also affected by the quality of the forcing datasets, especially precipitation. The model could compensate systematic biases of, e.g., net radiation or temperature by adjusting the modeled evaporative fraction or snowmelt factor respectively such that forcing biases are not propagated to fluxes but rather to these intermediate factors. Due to the water balance constraint, however, biases in the precipitation product cannot be compensated. Likewise, potential biases in the static input variables are no issue as the neural network exploits only patterns in the data irrespective of magnitudes or units.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

In the hybrid modeling framework, the quality of the observational constraints is also a source of uncertainty. While physically-based models heavily rely on detailed process descriptions, the hybrid model learns the responses from the data. Erroneous constraints will, thus, have an impact on the simulated hydrological responses and parameter estimations. While random errors can be counteracted by using more data, biases impact the model directly and cannot be mitigated. The data used in this study have well documented deficiencies: 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 model 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). However, the multi-objective optimization may dampen the negative effects of biases, as the model can trade off the different constraints and does not—and cannot due to physical constraints—fit the data perfectly.

Lastly, the model optimization process itself is a source of uncertainty. Therefore, the cross-validation splitting and neural network initialization were done randomly, and the model simulations were found to be relatively robust. In case of stability issues, which were rare (see Fig. B1), simple but "ad-hoc" approaches like gradient clipping or better initialization of the physical states were found to be sufficient. A better and systematic approach to understand the interplay of the neural network and the physically-based model, as well as the spin-up process, is needed.

4.3.2 Challenges and opportunities

The hydrological pathways in H2M are rather simple compared to GHMs, but the model still expresses a high data-adaptiveness as we demonstrated. While GHMs usually represent a wide range of subprocesses (e.g. infiltration, preferential flow, topographical runoff-runon), the hybrid model compiles them into only a few response functions and the model complexity and interactions within is, so to speak, outsourced to the neural network. Still, missing representations of storages (e.g., surface storage) and hydrological pathways (e.g., streamflows) limit the model flexibility and, to a certain extent, can corrupt the other latent variables as the model tries to accommodate for missing processes. At the same time, the relaxation of assumptions can be seen as an opportunity, as they may be wrong or incomplete.

As the model behavior emerges largely from the data, 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 do not allow to incorporate physical knowledge. GHMs themselves largely rely on prior knowledge, which may be wrong or incomplete and the model parameterization is usually not resolved regionally, resulting in model uncertainties (Beck et al., 2016) expressed in the disagreement of the model simulations. The hybrid model can be seen as a compromise between the purely data-riven and the physically-based approach, as physical principles (such mass conservation) are respected, but qualitative assumptions on the processes are still used. A great example to illustrate this are the unbound cumulative water deficit and groundwater pool. Although unlimited, the storages are constrained *implicitly* by the data and process descriptions: The partitioning of soil moisture and groundwater is partially achieved through data constraints (e.g.,

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

evapotranspiration), but also by the assumption that groundwater is a "slow" storage, achieved through a fractional baseflow. The partitioning is, thus, closely related to the decomposition of the total runoff into "slow" components (baseflow) and "fast" components (surface runoff, directly linked to rainfall and snowmelt).

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. These processes may be implicitly learned, but can also lead to biases in the simulations and parameter estimates. Snow sublimation, for example, plays an important role in the water balance but is difficult to parameterize (Bowling et al., 2004). The H2M model can compensate for snow sublimation by reducing snowfall or increasing snowmelt, which improves simulations of snow water equivalent, but introduces biases on snowfall, snowmelt, and the respective parameters. Similar problems arise from the missing representation of surface water storage and river routing, as previously discussed. Further, 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 constraints or better physical processes in order to avoid equifinality issues.

Equifinality occurs when multiple parameter combinations lead to the same or similar solutions. This is not an issue with the neural network, where equifinality comes by design, but with the parameter estimates and consequently with the hydrological responses, and has been reported for hydrological models (Beven and Freer, 2001). It is well possible that the decomposition into CWD and GW is not properly constrained under some circumstances, for example 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. The mathematical or conceptual framework to identify such equifinality issues is currently missing. More research is needed to address these problems, and, in particular, a complementary development of application-based models as done here, and smaller-scale, better constrained exercises to advance hybrid modeling can be a viable alternative.

One way to counteract equifinality is using additional data constraints. 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 extremely 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 to add further complexity to the model without increasing the uncertainty.

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 represent-

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

ing 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 the goal here 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).

Further opportunities lie on the representation of processes at different scales. In the presented hybrid model, we included static features with higher resolution than forcing variables to represent sub-grid scale processes and heterogeneity. A neural network compressed the dimensionality-reduced static variables before they were fed into the recurrent layer (a map of extracted features is shown in Fig. B2). Further work is needed to develop frameworks that do not only involve datasets by aggregating them into similar, easy-to-process chunks, but can efficiently integrate data at different spatial and temporal resolution. Finally, incorporating lateral interactions and flow between grid cells (e.g., large scale groundwater flow, 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 neighboring regions.

5 Conclusions

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

1. The hybrid model had similar performance as physically-based models on 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 follow 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 constraint.

3. The partitioning of the terrestrial water storage by the hybrid model 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 shows larger contribution by soil moisture.

4. Key opportunities and challenges in hybrid modeling to be addressed in 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 the knowledge-driven approaches. The data-adaptiveness can reveal weaknesses and strengths of process-based models and provide important insights for water cycle attribution and diagnostics. The findings of this study can be generalized to other spheres and scales, as long as sufficient data and process knowledge is available.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

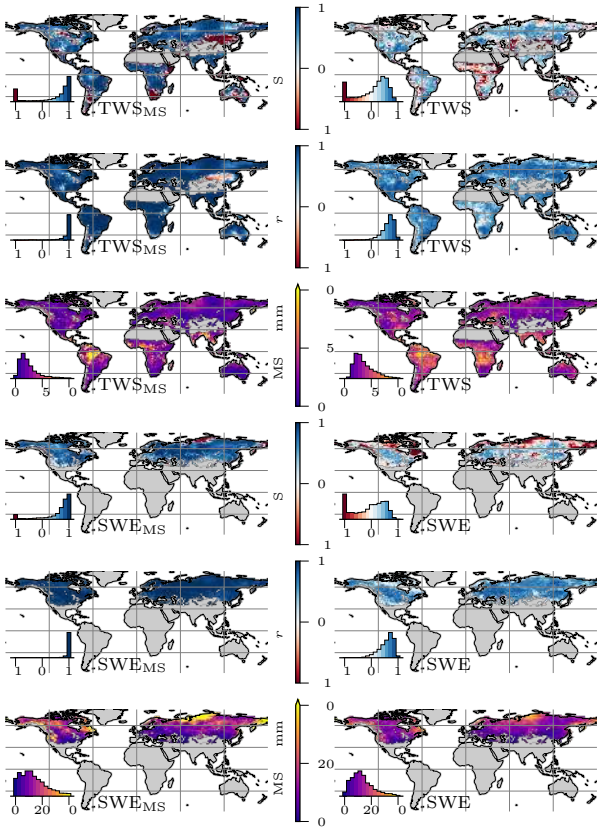


Figure A1. Local model performance for terrestrial water storage (TWS) and snow water equivalent (SWE) on the mean seasonal cycle (MSC) and the interannual variability (IAV) within the test period. The Nash–Sutcliffe model efficiency (NSE), Pearson correlation (r) and Root Mean Square Error (RMSE) are shown. Negative NSE have been remapped to a range -1 to 0 using the hyperbolic tangent function to avoid large negative values. The inset plots show the cell-area weighted histogram of the map values.

Code and data availability. The simulated hydrological data and the code are available here: https://dx.doi.org/10.17617/3.65. The code is also available on github: https://github.com/bask0/h2m. Note that we cannot share the the data used as model input, but all datasets are referenced in the manuscript.

Appendix A: Spatial model performance

Overall, high NSE of TWS_{MSC} is achieved in most regions (Fig. A1). Low TWS_{NSE} hotspots 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,

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions

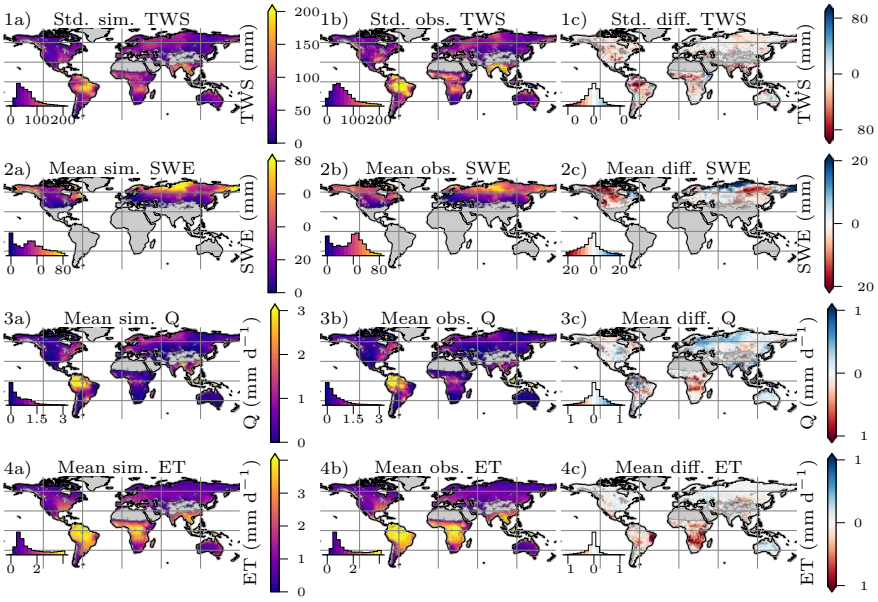


Figure A2. Mean a) simulated, b) observed, and c) difference of simulated - observed (positive means simulated is larger) terrestrial water storage (TWS, 1a–c), snow water equivalent (SWE, 2a–c), total runoff (Q, 3a–c), and evapotranspiration (ET, 4a–c). Note that for the TWS, the standard deviation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

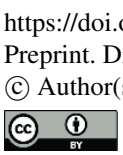
where the TWS variations are minimal. The TWS_{IAV} was also reproduced well in terms of r. The SWE_{MSC} is reproduced well in terms of NSE and r, while NSE for SWE_{IAV} is low especially in tundra regions (Fig. A1). The RMSE is also larger in high latitudes but more concentrated in regions with large 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 of 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 Central Africa, and slightly overestimated in in northwestern Eurasia, central Amazon and coastal regions of Australia and East Asia. ET, finally, is underestimated by the model, prominently in most of Subsaharan Africa and East Brazil, while no major biases are present in other regions.

Appendix B: Model optimization

The model optimization within the cross-validation setting is shown in Fig. B1. 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

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021



Hydrology and Earth System Sciences
Discussions
Open Access
EGU

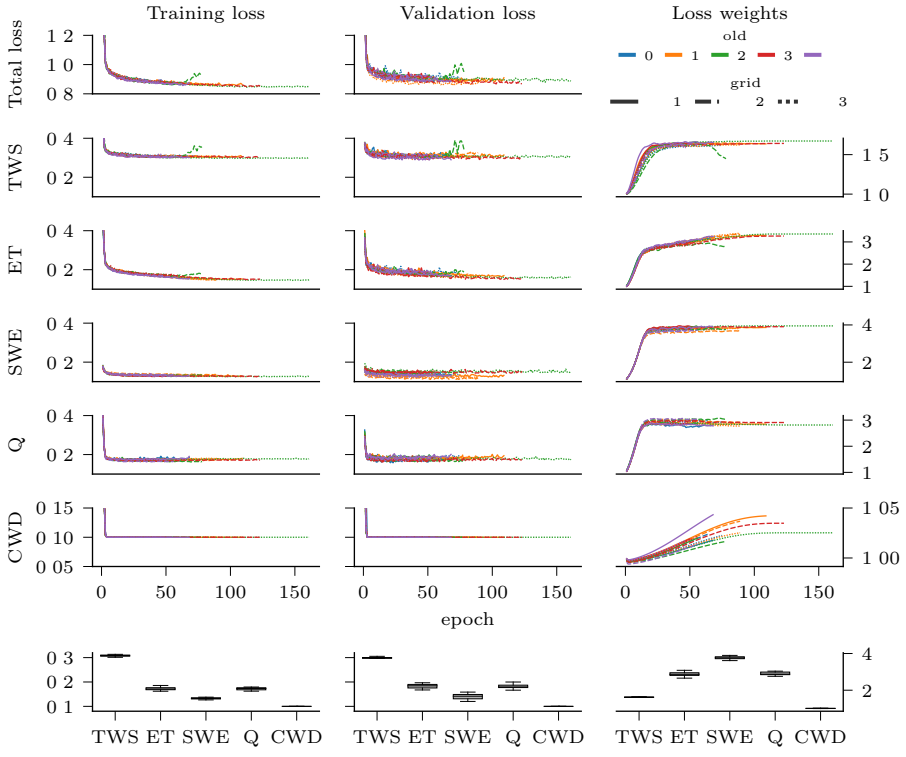


Figure B1. Model training process for the cross validation runs. The left and central column represent the unweighted total and variable-specific MSE loss. The right column shows how the variable weights developed over training time. The x-axis represents the number of iterations through the training 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.

early stopping mechanism, however, the model from the best validation loss is restored and used for the test set prediction. The loss and weight distributions at optimum across cross-validation runs were stable (bottom row of boxplots in Fig. B1). 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 larges 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 (variable-specific loss weights) were also 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 boxplots in Fig. B1, we can see that variables with a lower loss is 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.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hydrology and Earth System Sciences
Discussions

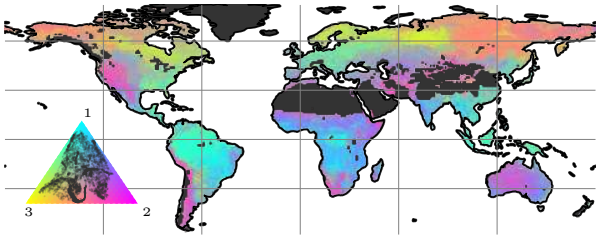


Figure B2. The t-distributed stochastic neighbor (t-SNE) reduction to 3 dimensions of static variable encoding (originally 12 dimensions) of one cross-validation run. The encoding is a low-level representation of the static inputs, i.e., soil and land-cover properties, learned by a neural network. The inset ternary plots show the distribution of the map values.

Part of the model tuning involved optimization of the sub-network FCNN^2 (Fig. 1), extracting features from the static variables which are then fed into the recurrent neural network. We visualized the outputs (*activations*) of the FCNN^2 to get an impression of the most relevant gradients within the static variables. For visualization, the twelve activations were reduced to three dimensions using t-SNE (Hinton and Roweis, 2002). The resulting map (Fig. B2) reveals patterns that seem very familiar: the component 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 hyper-parameters. Also, the reduction to three dimensions does only reveal the major gradients and does not represent the entire variability.

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 manuscript.

Competing interests. The authors declare that they have no conflict of interest.

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. In addition, we thank Uli Weber for data preprocessing and the colleagues from the MPI for Biogeochemistry for the inspiring discussions.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021


Hydrology and Earth System Sciences
Discussions

References

Andrew, R., Guan, H., and Batelaan, O.: Estimation of GRACE water storage components by temporal decomposition, Journal of Hydrology, 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., et al.: FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, Bulletin of the American Meteorological Society, 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 Resources Research, 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 Resources Research, 52, 3599–3622, https://doi.org/10.1002/2015WR018247, 2016.

Beck, H. E., van Dijk, A. I., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from ten state-of-the-art hydrological models, Hydrology and Earth System Sciences, 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., et al.: Status of high-latitude precipitation estimates from observations and reanalyses, Journal of Geophysical Research: Atmospheres, 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., pp. 443–476, Water Resources Publications, Colorado, USA, 1995.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of hydrology, 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

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

Bowling, L., Pomeroy, J., and Lettenmaier, D.: Parameterization of blowing-snow sublimation in a macroscale hydrology model, Journal of Hydrometeorology, 5, 745–762, https://doi.org/10.1175/1525-7541(2004)005<0745:POBSIA>2.0.CO;2, 2004.

Budyko, M. I.: Climate and life, vol. 18, Academic press, 1 edn., 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., et al.: Global land cover mapping at 30 m resolution: A POK-based operational approach, ISPRS Journal of Photogrammetry and Remote Sensing, 103, 7–27, https://doi.org/10.1016/j.isprsjprs.2014.09.002, 2015.

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, Journal of Hydrometeorology, 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, Journal of Geophysical Research: Atmospheres, 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.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

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.

Döll, P., Kaspar, F., and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, Journal of Hydrology, 270, 105–134, https://doi.org/10.1016/S0022-1694(02)00283-4, 2003.

Falkner, S., Klein, A., and Hutter, F.: BOHB: Robust and efficient hyperparameter optimization at scale, 2018.

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

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., and Otero-Casal, C.: Hydrologic regulation of plant rooting depth, Proceedings of the National Academy of Sciences, 114, 10 572–10 577, https://doi.org/10.1073/pnas.1712381114, 2017.

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

Fischer, C., Leimer, S., Roscher, C., Ravenek, J., de Kroon, H., Kreutziger, Y., Baade, J., Beßler, H., Eisenhauer, N., Weigelt, A., et al.: Plant species richness and functional groups have different effects on soil water content in a decade-long grassland experiment, Journal of Ecology, 107, 127–141, https://doi.org/10.1111/1365-2745.13046, 2019.

Geer, A.: Learning earth system models from observations: machine learning or data assimilation?, Philosophical Transactions of the Royal Society A, 379, 20200 089, https://doi.org/10.1098/rsta.2020.0089, 2021.

Getirana, A., Kumar, S., Girotto, M., and Rodell, M.: Rivers and floodplains as key components of global terrestrial water storage variability, Geophysical Research Letters, 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 System Science 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, 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., pp. 109–127, Springer, Dordrecht, The Netherlands, https://doi.org/10.1007/1-4020-4260-4_7, 2006.

Güntner, A.: Improvement of global hydrological models using GRACE data, Surveys in geophysics, 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 Resources Research, 43, 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., et al.: Multimodel estimate of the global terrestrial water balance: setup and first results, Journal of Hydrometeorology, 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, International journal of climatology, 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., et al.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, https://doi.org/10.1371/journal.pone.0169748, 2017.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

Hinton, G. and Roweis, S. T.: Stochastic neighbor embedding, in: NIPS, vol. 15, pp. 833–840, Citeseer, https://doi.org/10.5555/2968618.2968725, 2002.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 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, Surveys in Geophysics, 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, Agricultural and Forest Meteorology, 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., et al.: Belowground consequences of vegetation change and their treatment in models, Ecological applications, 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 Resources Research, 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., pp. 229–269, Cambridge University Press, https://doi.org/10.1017/CBO9781107415379.008, 2014.

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., et al.: 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, 2018.

Kim, H., Yeh, P. J.-F., Oki, T., and Kanae, S.: Role of rivers in the seasonal variations of terrestrial water storage over global basins, Geophysical Research Letters, 36, 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, Climate Dynamics, 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., et al.: Global distribution of groundwater-vegetation spatial covariation, Geophysical Research Letters, 44, 4134–4142, https://doi.org/10.1002/2017GL072885, 2017.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

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, The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 1537–1544, https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020, 2020.

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, Journal of Geophysical Research: Solid Earth, 124, 9332–9344, https://doi.org/10.1029/2019JB017415, 2019.

Larue, F., Royer, A., De Sève, D., Langlois, A., Roy, A., and Brucker, L.: Validation of GlobSnow-2 snow water equivalent over Eastern Canada, Remote sensing of environment, 194, 264–277, https://doi.org/10.1016/j.rse.2017.03.027, 2017.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I.: Tune: A research platform for distributed model selection and training, 2018.

Luojus, K., Pulliainen, J., Takala, M., Derksen, C., Rott, H., Nagler, T., Solberg, R., Wiesmann, A., Metsamaki, S., Malnes, E., et al.: 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, 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, Advances in Water Resources, 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, Advances in water resources, 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, Journal of hydrology, 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.

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-Pacific Journal of Atmospheric Sciences, pp. 1–14, https://doi.org/10.1007/s13143-019-00161-6, 2019.

Papagiannopoulou, C., Miralles, D. G., Demuzere, M., Verhoest, N. E., and Waegeman, W.: Global hydro-climatic biomes identified via multitask learning, Geoscientific Model Development, 11, 4139–4153, https://doi.org/10.5194/gmd-11-4139-2018, 2018.

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

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, Journal of Geophysical Research: Earth Surface, 112, https://doi.org/10.1029/2006JF000615, 2007.

Reichle, R. H.: Data assimilation methods in the Earth sciences, Advances in water resources, 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., et al.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195, https://doi.org/10.1038/s41586-019-0912-1, 2019.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

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., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoing, H., Lo, M., Müller-Schmied, H., Döll, P., et al.: Tracking seasonal fluctuations in land water storage using global models and GRACE satellites, Geophysical Research Letters, 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 Resources Research, 52, 9412–9429, https://doi.org/10.1002/2016WR019494, 2016.

Schellekens, J., Dutra, E., la Torre, A. M.-d., Balsamo, G., van Dijk, A., Weiland, F. S., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., et al.: A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset, Earth System Science Data, 9, 389–413, https://doi.org/10.5194/essd-2016-55, 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, Journal of 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-Science Reviews, 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., et al.: HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community, Hydrology and Earth System Sciences, 22, 5639–5656, https://doi.org/10.5194/hess-22-5639-2018, 2018.

Sperry, J. S. and Hacke, U. G.: Desert shrub water relations with respect to soil characteristics and plant functional type, Functional Ecology, 16, 367–378, https://doi.org/10.1046/j.1365-2435.2002.00628.x, 2002.

Sun, L., Seidou, O., Nistor, I., and Liu, K.: Review of the Kalman-type hydrological data assimilation, Hydrological Sciences Journal, 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 Resources Research, 44, 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, International Journal of Climatology, 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 Sensing of Environment, 115, 3517–3529, https://doi.org/10.1016/j.rse.2011.08.014, 2011.

Tootchi, A., Jost, A., and Ducharne, A.: Multi-source global wetland maps combining surface water imagery and groundwater constraints, Earth System Science 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., and et al.: 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.

https://doi.org/10.5194/hess-2021-211
Preprint. Discussion started: 25 May 2021

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, Hydrology and Earth System Sciences, 22, 4061–4082, https://doi.org/10.5194/hess-22-4061-2018, 2018.

Van Beek, L., Wada, Y., and Bierkens, M. F.: Global monthly water stress: 1. Water balance and water availability, Water Resources Research, 47, 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, International Journal of Geographical Information Science, 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., Renzullo, L., Wada, Y., Tregoning, P., et al.: A global water cycle reanalysis (2003–2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble, 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/doi.org/10.5065/PZ8F-F017, 2018.

Wada, Y., Wisser, D., and Bierkens, M. F.: Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, Earth System Dynamics Discussions, 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 Computing Surveys (CSUR), 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, Journal of Geophysical Research: Solid Earth, 120, 2648–2671, https://doi.org/10.1002/2014JB011547, 2015.

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, Bulletin of the American Meteorological Society, 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 Resources Research, 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 Resources Research, 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, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 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, Journal of climate, 21, 248–265, https://doi.org/10.1175/2007JCLI1639.1, 2008.