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.

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

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

The hybrid hydrological model (H2M), extended from Kraft et al. (2020), simulates the dynamics of snow, soil moisture, and groundwater storage globally at $1°$ spatial resolution and daily time step, where water fluxes are simulated by an embedded recurrent neural network. We trained the model simultaneously against observational products of terrestrial water storage variations (TWS), runoff (Q), evapotranspiration (ET), and snow water equivalent (SWE) with a multi-task learning approach. We find that the H2M is capable of reproducing key patterns of global water cycle components with model performances being at least on par with four state-of-the-art GHMs, which provides a necessary benchmark for H2M. The neural network learned hydrological responses of evapotranspiration and runoff generation to antecedent soil moisture states qualitatively consistent with our understanding and theory. Simulated contributions of groundwater, soil moisture, and snowpack variability to TWS variations are plausible and within the ranges of traditional GHMs. H2M identifies a somewhat stronger role of soil moisture for TWS variations in transitional and tropical regions compared to GHMs.

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

1 Introduction

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

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

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

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

One of the key hydrological data products for diagnosing and understanding global land water cycle variations is total terrestrial water storage (TWS). The TWS is an observation-based rasterized product that integrates all water storage components

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

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

We first describe the datasets used, the hybrid hydrological model (H2M), and the model training and evaluation approach in Section 2. We then present the benchmarking of H2M performance against a set of GHM simulations from the eartH2O ensemble, which are the measuring standard for H2M (Sect. 3.1). Section 3.2 provides the data-driven perspective on hydrological responses, followed by Sect. 3.3 that focuses on the TWS decomposition. Additional plausibility and interpretability of the H2M simulations are presented in Sect. 4.1 and Sect. 4.2. Lastly, we provide a more general assessment of the challenges and opportunities of the hybrid approach in Sect. 4.3.

2 Data and methods

2.1 Datasets

2.1.1 Meteorological forcing

Three time-varying meteorological datasets were used to force H2M: (Tab. 1).

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

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 storage variables and lower case for fluxes.

	Acr.	Math. notation	Spatial resolution		Temporal resolution	Dataset	Resources
			Native	Agg.			
Water cycle constraints							
Terrestrial water storage	TWS	T	0.50°	1.00°	Monthly	GRACE Tellus JPL RL06M v1	Watkins et al. (2015), Wiese et al. (2018)
Evapotranspiration	ET	e	0.50°	1.00°	Monthly	FLUXCOM v1	Tramontana et al. (2016), Jung et al. (2019)
Runoff	Q	q	0.50°	1.00°	Monthly	GRUN v1	Ghiggi et al. (2019)
Snow water equivalent	SWE	S	0.25°	1.00°	Daily	GlobSnow v2	Takala et al. (2011), Luojus et al. (2014)
Meteorological forcing							
Precipitation	-	p	1.00°	1.00°	Daily	GPCP 1dd v1.2	Huffman et al. (2012)
Net radiation	-	r_{net}	1.00°	1.00°	Daily	CERES SYN1deg Ed4A	Wielicki et al. (1996), Doelling (2017)
Air temperature	-	T_{air}	0.50°	1.00°	Daily	CRUNCEP v8	Harris et al. (2014), Viovy (2018)
Static variables							
Soil properties	-	-	1/120°	1/30°	-	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

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

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

95 To test the impact of the model forcings on the comparison with GHMs (Sect. 3.1.1), we carried out additional H2M simulation with forcing datasets from the Watch Forcing Data-ERA Interim (WFDEI) dataset (Weedon et al., 2014) in an independent setup (Appendix D).

2.1.2 Static variables

A set of temporally static variables was used to represent land surface characteristics (Tab. 1):

100 **i)** Soil properties from the soilgrids dataset (Hengl et al., 2017): *absolute depth to bedrock* and the average (along depth) of *bulk density*, *coarse fragments*, *clay*, *silt*, and *sand* (6 variables in total);

ii) Land cover fractions from the Globland30 dataset (Chen et al., 2015) for the 10 classes: *water bodies*, *wetlands*, *artificial surfaces*, *tundra*, *permanent snow and ice*, *grasslands*, *barren*, *cultivated land*, *shrublands*, and *forests*;

iii) Digital elevation model from GTOPO30 (DOI/USGS/EROS, 1997); and

iv) Fractions of groundwater-driven wetlands, regularly flooded wetlands, and the intersection of them (Tootchi et al., 2019), i.e., a total of 3 variables.

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

2.1.3 Observational constraints

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

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

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

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

iv) The daily SWE observations were obtained from the GlobSnow v2 product (Takala et al., 2011; Luojus et al., 2014). GlobSnow provides snow water equivalent in the Northern Hemisphere above 40°N, while the mostly snow-free Southern Hemisphere is not covered. In GlobSnow, the time steps with no snow are encoded as missing values. Thus, we gap-filled the

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

| | | −CWD (fast storage) | | GW (slow storage) | |
	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 interception, SStor=surface storage

SWE observations using 8 d snow cover fraction (SCF) from MODIS (Hall and Riggs, 2016), disaggregated to daily resolution using nearest neighbor method, to obtain a global coverage, albeit with 0 SWE. To do so, a time step in the SWE was set to 0 if the grid-level SWE a) SCF $\pm$ 12 d was in average below 10 % and b) all SWE observations SCF $\pm$ 12 d were missing.

2.1.4 Global hydrological model ensemble

To evaluate the H2M simulations of TWS and its components, we selected the GHMs from the eartH2O ensemble (Schellekens et al., 2017), version WWR1. From the ten available model simulations, we selected those including groundwater storage: LIS-FLOOD (Van Der Knijff et al., 2010), W3RA (Van Dijk and Warren, 2010; Van Dijk et al., 2014), PCR-GLOBWB (Van Beek et al., 2011; Wada et al., 2014), and SURFEX-TRIP (Decharme et al., 2010, 2013).

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

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

2.1.5 Data filtering

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

1. Grid cells with more than 50 % water bodies, more than 90 % permanent snow or ice, or more than 90 % bare land.

2. Regions with more than 90 % artificial built-up surfaces.

3. Regions with large groundwater withdrawals labeled as "Groundwater depletion" under athropogenic influence in Rodell et al. (2018).

4. Grid cells with more than 50 % missing values in any of the time series of the observational constraints.

5. Mountainous areas which are masked in GlobSnow.

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

2.2 The hybrid hydrological model (H2M)

The H2M consists of a dynamic neural network and a simple hydrological framework that represent the major water fluxes and changes in water storage (Fig. 1). The H2M is set up as a "global" model, i.e., the same model is used to predict the full spatio-temporal domain, in contrast to separate models for each grid cell in the "local" setup. The H2M only considers the vertical flow/transport of the water through the system and does not include the lateral flow of either surface (river routing) or sub-surface water (groundwater flow).

The neural network (Sect. 2.2.2) yields a set of time-varying coefficients conditioned on the meteorological forcing and spatial properties derived from the static input variables. These coefficients (e.g., snowmelt factor) are then used in a set of hydrological equations that are introduced in Sect. 2.2.1.

For the sake of consistency and clarity, α denotes the time-varying coefficients that are directly estimated by the neural network, and β denotes the global parameters that are learned as spatially constant. Throughout the manuscript, t is used as time index and i as the grid cell index. Uppercase variables are used for physical state variables.

2.2.1 Hydrological components

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

Snow

Snow water equivalent is one of the water storages simulated by the H2M model, and it is also constrained by the corresponding observation during model training.
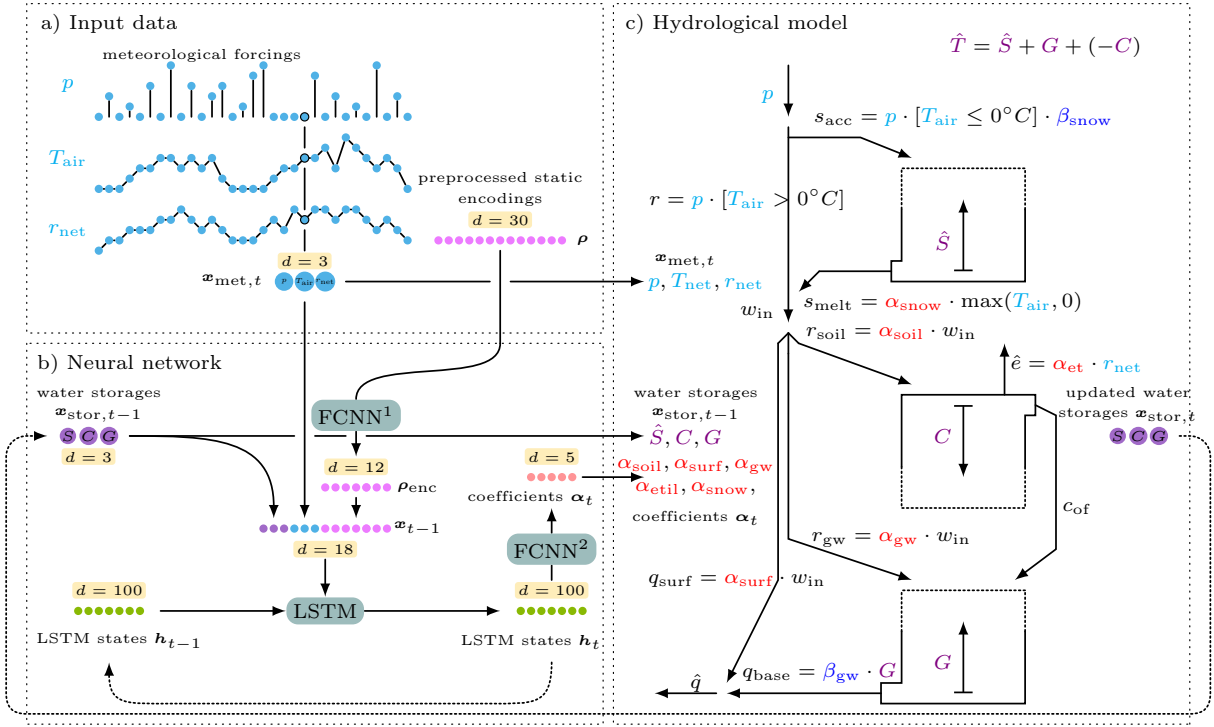
Figure 1. The H2M dynamically updates the water storages $\boldsymbol{x}_{\text{stor},t-1}$ conditioned on meteorological $\boldsymbol{x}_{\text{met},t}$ and static inputs $\boldsymbol{\rho}$. The neural network module (b) receives the inputs (a) and the antecedent water storages and yields a set of physical coefficients $\boldsymbol{\alpha}_t$ used in the hydrological module (c). The neural network contains a long short-term memory (LSTM) layer maintaining its own internal state $\boldsymbol{h}$ (cell state omitted here), and two fully connected networks (FCNN). Solid arrows and densely dotted lines denote value transfer and recurrent connections, respectively. The hat operator $(\hat{\cdot})$ denotes the variables that are constrained with observations, and upper case variables are storages. Forcings (cyan): p: precipitation, T_{air}: air temperature, r_{net}: net radiation. Water storages (purple): $\hat{S}$: snow water equivalent, C: cumulative soil water deficit, G: groundwater, $\hat{T}$: terrestrial water storage. Time-varying coefficients (red): α_{soil}: soil recharge fraction, α_{gw}: groundwater recharge fraction, α_{surf}: surface runoff fraction, α_{smelt}: snowmelt coefficient fraction, α_{et}: evaporative fraction. Learned global constants (blue): β_{snow}: snow undercatch correction constant, β_{gw}: baseflow constant. Water fluxes: r: rainfall, s_{acc}: snow accumulation, s_{melt}: snowmelt, w_{in}: liquid phase water inout, r_{soil}: soil recharge, r_{gw}: groundwater recharge, q_{surf}: surface runoff, q_{base}: baseflow, $\hat{e}$: evapotranspiration, c_{of}: overflow, $\hat{q}$: total runoff.

Snow accumulation

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

is precipitation with air temperatures $T_{\text{air}} \leq 0°\text{C}$. The accumulation is scaled by a learned (optimized) global constant $0 <$
185    $\beta_{\text{snow}} < 1$. The correction accounts for the known overestimation of solid precipitation due to over-correction for under catch

of snowfall in gauge measurements (Decharme and Douville, 2006). Potential snowmelt

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

is then calculated using a degree-day approach. Opposite to snow accumulation, s_{melt} occurs under the condition of $T_{\text{air}} > 0°\text{C}$. The time-varying snowmelt coefficient α_{smelt} is estimated by the neural network and mapped to positive values by applying the softplus activation function

$$\text{softplus}(x) = \log(1 + \exp(x)) \quad . \qquad (3)$$

The snow water equivalent

$$S_{t,i} = \max(S_{t-1,i} + s_{\text{acc},t,i} - s_{\text{melt},t,i}, 0) \qquad \text{(in mm)} \qquad (4)$$

is then updated using snow accumulation and melt. Positive values of S are enforced by truncating negative values using the maximum function.

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

Soil recharge, groundwater recharge, and surface runoff

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

The parameters for the partitioning are estimated by the neural network and mapped to the range $(0,1)$, and naturally constrained to the sum of 1 by applying the softmax transformation (Goodfellow et al., 2016). A softmax transformation generalizes the logistic function to multiple dimensions. Note that the constrained training of parameters to 1 ensures that the incoming water is neither lost or generated during the partitioning respecting the physical law for the conservation of mass.

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

$$r_{\text{soil},t,i} = \alpha_{\text{soil},t,i} \cdot w_{\text{in},t,i} \qquad \text{(in mm d}^{-1}\text{)}, \qquad (5)$$
$$r_{\text{gw},t,i} = \alpha_{\text{gw},t,i} \cdot w_{\text{in},t,i} \qquad \text{(in mm d}^{-1}\text{)}, \text{ and} \qquad (6)$$
$$q_{\text{surf},t,i} = \alpha_{\text{surf},t,i} \cdot w_{\text{in},t,i} \qquad \text{(in mm d}^{-1}\text{)}, \qquad (7)$$

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

Evapotranspiration and soil moisture

The total evapotranspiration

$$e_{t,i} = \alpha_{\text{et},t,i} \cdot \frac{r_{\text{net},t,i}}{2.45} \qquad \text{(in mm d}^{-1}) \tag{8}$$

is calculated as the product of the evaporative fraction α_{et} and net radiation r_{net} (MJ d^{-1} m^{-2}) converted to mm d^{-1} assuming a latent heat of vaporization of 2.45 MJ mm^{-1} m^{-2}. The evaporative fraction is learned by the neural network and mapped to the range $(0, 1)$ by applying the sigmoid activation function. Note that evapotranspiration is constrained by the corresponding observation during model training.

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

$$C^*_{t,i} = C_{t-1,i} + r_{\text{soil},t,i} - e_{t,i} \qquad \text{(in mm)}, \tag{9}$$

$$c_{\text{of},t,i} = \log(1 + \exp(-C^*_{t,i})) \qquad \text{(in mm d}^{-1}\text{), and} \tag{10}$$

$$C_{t,i} = C^*_{t,i} - c_{\text{of},t,i} \qquad \text{(in mm)}, \tag{11}$$

which has the benefit of having a physical saturation limit of 0. For the comparison with the GHMs (Sect. 3.1.1), we calculate soil moisture (mm) dynamics as $M = -C$. The state C is updated by addition of the soil recharge r_{soil}, subtraction of evapotranspiration e (Eq. 9), and leveling by the overflow mechanism (Eq. 10–11): If C approaches 0, an overflow mechanism allows for direct discharge of excess soil moisture into the deeper groundwater storage. Due to the heterogeneity within a model cell, the overflow c_{of} starts already at values close to 0, which is achieved by using the softplus function (Eq. 10).

Baseflow and Groundwater

The baseflow

$$q_{\text{base},t,i} = G_{t-1,i} \cdot \beta_{\text{gw}} \qquad \text{(in mm d}^{-1}) \tag{12}$$

is calculated as fraction of the past groundwater storage G_{t-1} and the global baseflow constant β_{gw}. Once the baseflow, groundwater recharge, and overflow of soil storage are calculated, the groundwater storage

$$G_{t,i} = G_{t-1,i} + c_{\text{of},t,i} + r_{\text{gw},t,i} - q_{\text{base},t,i} \qquad \text{(in mm)} \tag{13}$$

can be updated using a simple water balance. In H2M, G represents an unconfined aquifer with an unlimited storage capacity.

Total Runoff

The total runoff

$$q_{t,i} = q_{\text{surf},t,i} + q_{\text{base},t,i} \qquad \text{(in mm d}^{-1}) \tag{14}$$

is simply calculated as the sum of the surface runoff q_{surf} (Eq. 7) and the baseflow q_{base} (Eq. 12). We emphasize here that the neural network receives the state of water storage as inputs and is, thus, able to learn interactions of the water storages, the input variables, and the corresponding hydrological partitioning and outflow coefficients. Thus, the runoff generation and evapotranspiration processes do not only depend on the current and past meteorological condition and static variables, but also on hydrological state, e.g., the soil water deficit. Therefore, we additionally use runoff as a data constraint during model training.

H2M storage components

For model training against GRACE, the variations of modeled terrestrial water storage components are added to calculate the total terrestrial water storage

$$T_{t,i}^* = S_{t,i} + G_{t,i} + (-C_{t,i}) \hspace{5cm} \text{(in mm)}. \hspace{2cm} (15)$$

Note that $-C$ is used in Eq. 15 as C itself is defined as the water *deficit*. As the observations of the terrestrial water storage from GRACE represent the temporal *variations*, the mean of simulated storage were removed from each grid cell as

$$T_{t,i} = T_{t,i}^* - \frac{1}{\mathcal{T}} \cdot \sum_{k=1}^{\mathcal{T}} T_{k,i}^* \hspace{5cm} \text{(in mm)}, \hspace{2cm} (16)$$

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

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

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

GW storage represents all delayed residual liquid water storage with infinite capacity. It is constrained by the baseflow fraction and subsequently temporal variation of total runoff (Eq. 12), which leads to a delayed dynamics compared to CWD.

2.2.2 The neural network (NN) module

The NN module (Fig. 1b) consists of three consecutively arranged sub-modules employed for extractions of different features. Overall, the NN module learns spatio-temporally varying coefficients of the hydrological model using meteorological and dimensionality-reduced static variables. The pseudo-code of the NN module is presented in Appendix E, while the sub-modules are introduced here.

The first feed-forward (i.e., non-temporal) sub-module learns a compressed representation of the static variables (Eq. 17). This representation, together with meteorological input, is then fed into the second sub-module, a recursive long short-term

memory (LSTM) model (Hochreiter and Schmidhuber, 1997), shown in Eq. 18. The third sub-module (Eq. 19) transforms the outputs of the LSTM to a set of coefficients, which are then fed into the hydrological components. As the model weights are shared across all grid cells, the NN module learns from the global dynamics and not exclusively from each grid cell. For a comprehensive overview of the neural network architectures, see Goodfellow et al. (2016), available online at www. deeplearningbook.org.

The first sub-module

$$\boldsymbol{\rho}_{\mathrm{enc},i} = f_{\mathrm{FCNN}^1}(\boldsymbol{\rho}_i) \tag{17}$$

is a fully-connected neural network (FCNN[1] in Fig. 1) with a single hidden layer and 150 nodes. It takes the static encodings $\boldsymbol{\rho}$ (see Sect. 2.1.2) as inputs and transforms them non-linearly into a more condensed form ($\boldsymbol{\rho}_{\mathrm{enc}}$). This reduces the high dimensionality of static inputs from 30 to 12 values. Ideally, this lower-dimensional representation describes the most significant gradients of the land characteristics at the sub-grid scale (visualized in Fig. C2, Appendix C). Note that the static variables have already been compressed in a pre-processing step, and the transformation in this sub-module is optimized specifically for the parameterization of the hydrological components.

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

$$\boldsymbol{h}_t = f_{\mathrm{RNN}}(\boldsymbol{h}_{t-1}, \boldsymbol{x}_{t,i}) \tag{18}$$

is updated at each time step by using interactions of the previous states $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{x}_{t,i}$. In H2M, $\boldsymbol{x}_{t,i}$ is a multivariate input consisting of concatenated current meteorological conditions $\boldsymbol{x}_{\mathrm{met},t,i}$, antecedent physical states from the hydrological model $\boldsymbol{x}_{\mathrm{stor},t-1,i}$, and the static features $\boldsymbol{\rho}_{\mathrm{enc},i}$ from Eq. 17. The input allows the LSTM to learn interactions among the variables conditioned on static land properties like land cover type or elevation. In the optimization process, the RNN learns to maintain a memory of information from past time steps and is capable of updating, removing, and extracting information.

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

Lastly, the third sub-module

$$\boldsymbol{\alpha}_{t,i} = f_{\mathrm{FCNN}^2}(\boldsymbol{h}_t) \tag{19}$$

linearly maps the LSTM output $\boldsymbol{h}_t$ to the coefficients $\boldsymbol{\alpha}$ of the hydrological components (FCNN[2] in Fig. 1). The vector $\boldsymbol{\alpha}$ contains five time-varying scalars corresponding to soil recharge fraction α_{soil}, groundwater recharge fraction α_{gw}, surface runoff fraction α_{surf}, snowmelt coefficient α_{smelt}, and evaporative fraction α_{et}.

2.3 Model training

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

2.3.1 Cross-validation setup

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

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

The hyper-parameters of the NN (i.e., the number of layers and hidden nodes in the neural networks, the learning rate, weight decay, dropout, and gradient clipping) are determined on a single grid, and the cross-validation is only applied on the remaining three grids. For 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).

This setup was chosen to avoid over-fitting, which is needed due to the data adaptivity of neural networks. In addition to the spatial and temporal splitting and the early stopping, we used weight decay (Loshchilov and Hutter, 2017) for regularization.

2.3.2 Training setup

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

Following a common practice in machine learning, the input variables and the observational data constraints are each z-transformed individually to follow a standard normal distribution using the pre-computed mean and standard deviations from

the training set. For physical consistency, the corresponding non-transformed variables are used for the hydrological balance equations (see Sect. 2.2).

340   To obtain an equilibrium of physical and hidden states of H2M, a model spin up is carried out with spin up data of five years duration, with each full year selected randomly from the training set. In each optimization iteration, the model is first forced by the spinup data to retrieve steady states, which are then used as initial conditions during the full forward run.

2.3.3 Multi-task loss

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

345   1) The loss term

$$\mathcal{L}_v(f_{\boldsymbol{\phi},\boldsymbol{\beta}}, \boldsymbol{x}, \boldsymbol{y}) = \sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{\mathcal{I}} ||y_{v,t,i} - \hat{y}_{v,t,i}||^2 \quad , v \in \{T, S, e, q\} \tag{20}$$

is calculated as the weighted sum of squared residuals for each (z-transformed) observational data constraint. Here, $y_{v,t,i}$ and $\hat{y}_{v,t,i}$ are the observed and predicted values of the variable v, respectively. An additional loss term is employed for regularization to promote parameters that would lead to near zero cumulative soil water deficit C (soil becomes saturated) at least

350   occasionally:

$$\mathcal{L}_{v=C}(f_{\boldsymbol{\phi},\boldsymbol{\beta}}, \boldsymbol{x}) = \sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{\mathcal{I}} (p_{10}(\hat{C}_{t,i}) + b_{\text{c}}) \cdot w_{\text{c}} \quad . \tag{21}$$

This term pushes the lower 10 percentile p_{10} of C towards zero. It was needed to reduce the state drift mostly related to spinup with random years of data that resulted in non-interepretable offsets in C (Kraft et al., 2020). A bias $b_{\text{c}} = 0.1$ was added to prevent the loss from becoming zero, which would interfere with the multitask loss weighting described below. The loss weight

355   w_{c} was lowered consecutively during training such that the loss $\mathcal{L}_{\text{C}}$ had only an impact during the early training phase.

2) A task uncertainty term $\boldsymbol{\sigma}$, weighting the individual losses dynamically:

$$\mathcal{L}_{\text{total}}(f_{\boldsymbol{\phi},\boldsymbol{\beta}}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) = \sum_{v \in \{T, S, e, q, C\}} \frac{1}{2 \cdot \sigma_v^2} \mathcal{L}_v + \log(\sigma_v) \quad , \tag{22}$$

where $\boldsymbol{\sigma}$ is a vector of task-specific uncertainties used to give more or less weight to a particular loss term. The task-specific uncertainties are trained during optimization so that the emphasis on a specific task changes dynamically over the course of the

360   model optimization. Note that $\log(\sigma_v)$ prevents the weights from diverging to infinity. This approach, called *self-paced multi-task weighting* (Kendall et al., 2018), is advantageous as the weights do not need to be subjectively predefined. The weights are visualized in Fig. C1, Appendix C.

Hence, the global optimization problem can be expressed as

$$\boldsymbol{\theta}^* = \operatorname*{arg\,min}_{\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\sigma})} \mathcal{L}_{\text{total}}(f_{\boldsymbol{\phi},\boldsymbol{\beta}}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) \quad , \tag{23}$$

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

2.4 Model evaluation and analysis

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

2.4.1 Performance metrics

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

$$e_{\text{NSE}} = 1 - \frac{\sum_{i=1}^{N}(m_i - o_i)^2}{\sum_{i=1}^{N}(o_i - \bar{o})^2} \quad , \tag{24}$$

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

2.4.2 Temporal and spatial scales

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

$$V_{\text{MSC},m} = \frac{1}{Y} \sum_{y=1}^{Y} V_{m,y} \quad , \text{ and} \tag{25}$$

$$V_{\text{IAV},m,y} = V_{m,y} - V_{\text{MSC},m} \quad , \tag{26}$$

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

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

For the global scale performance we calculate the metrics in two different ways that produce a single metric at the global scale by a mapping function $f_{\text{perf}} : \mathbb{R}^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}} \mapsto \mathbb{R}$ that compares two sequences of length $\mathcal{T}$. The first, which we call the *global performance*

$$\mathcal{M}_{\text{global}} = f_{\text{perf}}\left(\left\{\text{mean}_{\text{spatial}}(\{m_{t,i}\}_{i=1,\dots,\mathcal{I}}), \text{mean}_{\text{spatial}}(\{o_{t,i}\}_{i=1,\dots,\mathcal{I}})\right\}_{t=1,\dots,\mathcal{T}}\right) \tag{27}$$
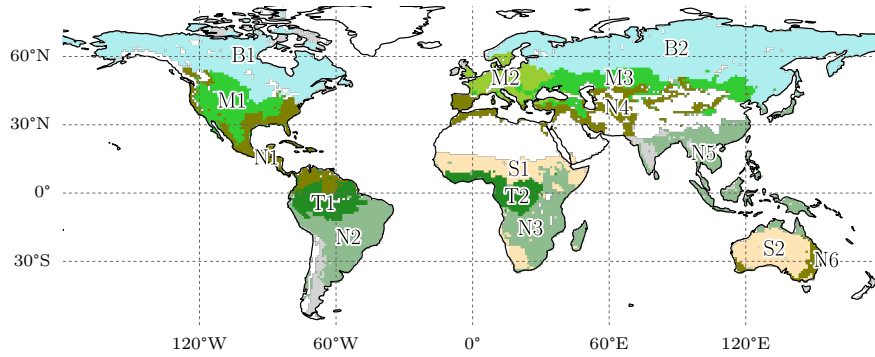
Figure 2. Continental hydro-climatic regions, adapted from Papagiannopoulou et al. (2018). **Boreal**: North America (B1) and Eurasia (B2). **Temperate**: North America (M1), Europe (M2), and Asia (M3). **Transitional**: North and Central America (N1), South America (S2), Africa (N3), Eurasia and North Africa (N4), Southeast Asia (N5), and Australia (N6). **Subtropical**: Africa (S1) and Australia (S2). **Tropical**: South America (T1) and Africa (T2). **Change:** Increased figure size.

395   represents the performance of the globally-aggregated variables. Similar to regional scale evaluations, these metrics reflect how the area-weighted globally aggregated time-series of observation and models compare. The global scale signal are themselves useful indicators, as they are often used to characterize the Earth system and land surface processes, e.g., climatic changes (Pachauri et al., 2014), or to evaluate water-carbon relations (Jung et al., 2017; Humphrey et al., 2016).

In contrast, global summary of the *local performance*

$$
\mathcal{M}_{\text{local}} = \text{median}_{\text{spatial}}\left(\left\{f_{\text{perf}}(m_{t,i}, o_{t,i})\right\}_{i=1,\ldots,\mathcal{I}}\right) \tag{28}
$$

is indicative of how the model performs locally all over the world. The local metrics are useful because the positive and negative model errors and tendencies can compensate when aggregated over a large spatial extent (e.g., Jung et al., 2017).

2.4.3   Terrestrial water storage variations and decomposition

For the analysis on the decomposition of TWS (Sect. 3.3 and Sect. 4.2.2), we use the simulated variables SWE, GW, and CWD
405   to assess their contributions to the TWS dynamics, seasonality, and interannual variability. Note that CWD represents a *deficit* of water in the soil. As a consequence, CWD shows opposite dynamics to water storages.

We calculate the absolute

$$
\mathcal{A}_v = \sum_{t=1}^{\mathcal{T}} |V_{v,t} - \bar{V}_v|, \quad v \in \{-C, G, S\} \tag{29}
$$

and relative contribution (hereinafter simply *contribution*)

410   $$
\mathcal{C}_v = \frac{\mathcal{A}_v}{\sum_{w \in \{-C,G,S\}} \mathcal{A}_w} \quad v \in \{-C, G, S\} \tag{30}
$$

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

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

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

3   Results

We first assess the performance of H2M simulations against the four observational data constraints (TWS, SWE, Q, and ET) at different spatial and temporal scales. This is followed by a comparison and benchmarking of model performance of H2M TWS and SWE against the simulations from four GHMs in the eartH2O ensemble. As the hybrid modeling framework has been significantly developed since Kraft et al. (2020), the H2M performance needs to be re-evaluated here. After the evaluations, we present a closer analysis and interpretation of the parameters estimated by the neural network that define the hydrological responses and generation of key hydrological fluxes in H2M. Finally, we present and compare the partitioning of TWS components.

3.1   General model performance

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

The model reproduced the patterns of the observed variables well (Tab. 3). In general, the global signal (global performance, see Eq. 27) was reproduced better than the local cell-level signal (median grid cell performance, see Eq. 28). For both observational constraint variables TWS and SWE an NSE $e_{\text{NSE}} > 0.8$ and $r > 0.9$ for the global signal and a cell median of NSE $e_{\text{NSE}} > 0.5$ and $r > 0.8$ was achieved. The seasonal signals of TWS_{MSC} and SWE_{MSC} were modeled with high accuracy ($e_{\text{NSE}} > 0.9$ on global, $e_{\text{NSE}} = 0.7$ on local level) while the interannual variability performance varied: The TWS_{IAV} was reproduced well with NSE $e_{\text{NSE}} = 0.54$ ($r = 0.8$) on global, and with NSE $e_{\text{NSE}} = 0.26$ ($r = 0.67$) on local level. The SWE_{IAV} performance was decent for the global signal ($e_{\text{NSE}} = 0.22$, $r = 0.87$), but lower ($e_{\text{NSE}} = 0.15$, $r = 0.64$) on local level.

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

3.1.1 Benchmarking H2M against GHMs

For the quantitative benchmarking of H2M performance with the state-of-the-art GHMs from eartH2O (see Sect. 2.1.4), we use the common time period of 2009 to 2012 (not 2009-2014 as in the previous section) but all common grid cells between the GHMs and H2M. This is justified as H2M has a negligible generalization error in space, i.e., the H2M performance is not systematically better in training grid cells. Similarly, we use the entire common time period (including the training data) for the *qualitative* assessment of the water cycle dynamics, as also in time, the generalization error was small. We note here that H2M was optimized with the datasets used for evaluation here, while the GHMs have either been calibrated using catchment-level observational runoff data (LISFLOOD) or rely on prior parameter estimation (W3RA, SURFEX-TRIP, PCR-GLOBWB) alone (Schellekens et al., 2017). The comparison presented here serves the purpose of performance benchmarking of hybrid modeling approach rather than finding the "best" model.

The H2M modeling efficiency falls within the range of the GHMs in terms of the global performance ($\diamond$ in Figure 3), although the performance varies less across the variables and temporal scales. However, H2M achieves a consistently higher local performance (boxes in Figure 3). The TWS is reproduced slightly better by the PCR-GLOBWB, which, however, has a relatively low performance on the local scale. All models struggle to reproduce the SWE_{IAV} signal: The median NSE of H2M is on a par with W3RA and SURFEX-TRIP, while the performance on spatially aggregated level is lower. A comparison of the model performance using the same forcings as in the eartH2O ensemble is provided in Appendix D, Fig. D1.
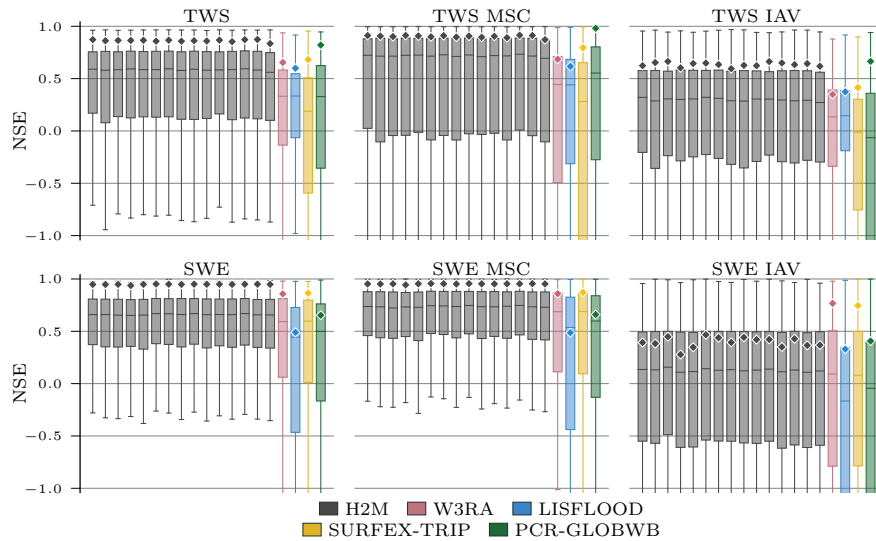
Figure 3. Global and local grid cell–level Nash–Sutcliffe model efficiency coefficient (NSE) of the hybrid hydrological model (H2M) and the process-based global hydrological models (GHMs) for the terrestrial water storage (TWS) on top and the snow water equivalent (SWE) at the bottom. The gray bars represent individual cross-validation runs. The ⋄-markers show the global (spatially averaged signal) model performance, the boxes represent the spatial variability of the local cell-level performance. The y-axis was cut at -1 due to some large negative NSE values. The panels show the model performance 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 2009 to 2012 on monthly time scale. Note that deviations from the numbers reported in Tab. 3 are due to different time ranges.

While all models reproduce the global monthly and seasonal TWS (Fig. 4) relatively well, the results vary more substantially for the TWS_{IAV}. Here, the H2M, WR3A, and LISFLOOD models show the best agreement with the TWS observations (also see Fig. 3 of model performance). The lower agreement of SURFEX-TRIP and PCR-GLOBWB on the global interannual

460 scale can be attributed to the time periods 2005–2006 and 2008–2010, respectively. From Fig. B1 of the regional averages (Appendix B), it becomes evident that this low agreement on global level can be attributed mainly to the tropical regions (T1: S-AM tropical and T2: AFR tropical).

The global SWE was well reproduced by H2M, especially the seasonal cycle showed better agreement than the GHMs, where the latter agreed well with the timing, but not the magnitude (Fig. 5). The global interannual variability was not reproduced

465 well by the H2M, LISFLOOD, and PCR-GLOBWB, all with an NSE below 0. Interestingly, H2M performed the best when forced by the same WFDEI forcing as in the GHM simulations (Fig. D1 in Appendix D). Regional model comparison of the time series are provided in Fig. B1 and B2, Appendix B.
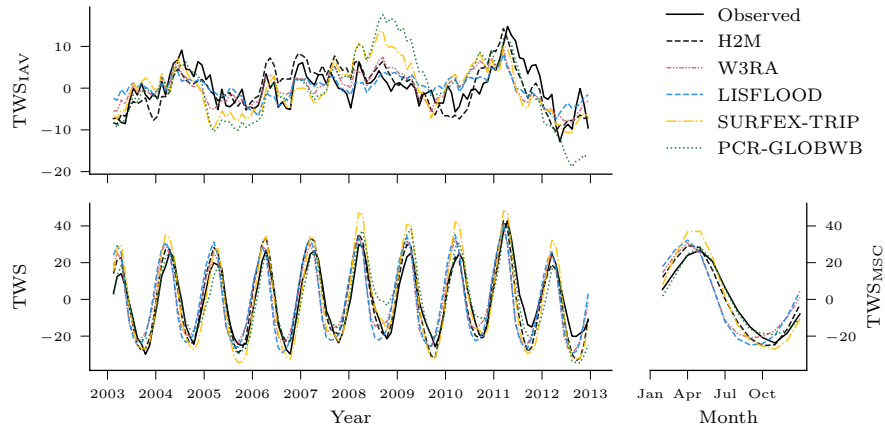
Figure 4. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the terrestrial water storage (TWS), its mean seasonal cycle (TWS$_{MSC}$) and its interannual variability (TWS$_{IAV}$) in mm for the global signal. The time series were aggregated using the cell size weighted mean across all grid cells. The regional time series are show in Appendix B, Fig. B1.
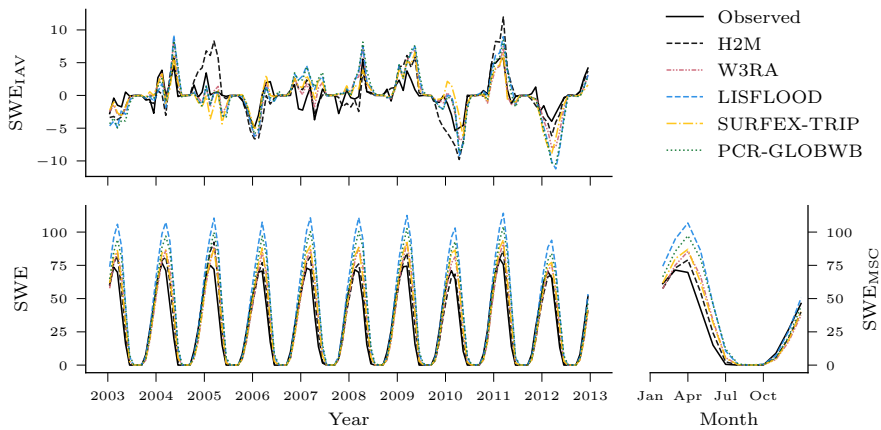
Figure 5. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the snow water equivalent (SWE), its mean seasonal cycle (SWE$_{MSC}$) and its interannual variability (SWE$_{IAV}$) in mm for the global signal. The time series were aggregated using the cell size weighted mean across all grid cells. The regional time series are show in Appendix B, Fig. B2.

3.2 Hydrological responses in H2M

For the qualitative assessment of the hydrological responses, we use all grid cells, like in the previous section, and show
470 the time range from 2003 to 2014 in time series plots. This involves the training data, but the impact is minimal due to a negligible generalization error. The H2M yields a set of data-driven, spatio-temporally varying coefficients that define the
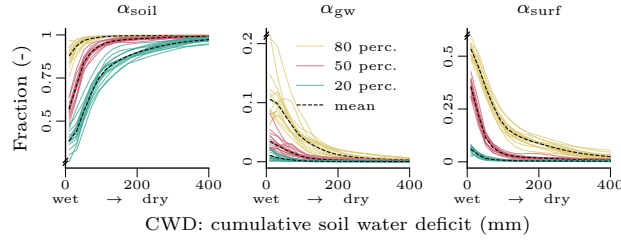
Figure 6. Relationship between the water input partitioning fractions for soil (α_{soil}, left), groundwater (α_{gw}, middle), and surface runoff (α_{surf}, right), and the cumulative soil water deficit (CWD) as learned by the neural network. The figure shows the respective percentiles of the spatio-temporal conditional distribution $P(\alpha \mid C \in B_i)$, where C is the cumulative soil water deficit on the x-axis discretized into $N = 10$ bins $B = \big\{ [0, 40), \ldots, [360, 400) \big\}_{i=1,\ldots,N}$. The colored lines show the percentiles per cross-validation run, the black dashed lines show the mean across the colored lines. The CWD dynamics correspond to negative soil moisture, i.e., larger CWD for dryer soils, and thus a larger CWD corresponds to smaller soil moisture. The plots are based on global daily cell time steps from 2009 to 2014. Note the differences in y-scale.

hydrological responses and generation of key hydrological fluxes. In particular, we focus on four parameters: α_{soil}, the fraction of throughfall that percolates into the soil; α_{gw}, the fraction that recharges the groundwater, α_{surf}; the fraction that runs off as surface runoff component; and α_{et}, the evaporative fraction (ratio of evapotranspiration to net radiation). Here, we analyze the

475 spatio-temporal variability of the parameters and how they are associated with the antecedent moisture condition defined by soil water deficit. In essence, these are analogous to stage-discharge relationships (Kumar, 2011) that are commonly used to characterize hydrological responses of river discharge at the catchment scale.

The partitioning of the liquid water input w_{inp} (rainfall plus snowmelt) using the fractions for soil recharge (α_{soil}), groundwater recharge (α_{gw}), and surface runoff (α_{surf}) was robust across cross-validation runs and showed a clear relationship to

480 CWD (Fig. 6). With an increasing soil water deficit (larger CWD, dryer soil), the soil recharge increases, while the groundwater recharge and surface runoff decrease. For a CWD below 200 mm, we observe a large spatio-temporal variation in the partitioning, evident through the relatively large difference between the 20$^{\text{th}}$ and 80$^{\text{th}}$ percentiles. The transition from larger soil recharge to larger groundwater recharge and surface runoff is exponentially decreasing, i.e., the change is faster with lower CWD (wetter soil). Above a CWD of 200 mm (dry soil), the partitioning is constant in space and time with α_{soil} converging

485 to 1, while α_{gw} and α_{surf} converge to 0.

In most hydroclimatic regions, α_{et} showed a negative relationship to CWD under dry conditions (large CWD), and no relationship in presence of precipitation or snowmelt (Fig. 7). The high latitude and tropical regions showed a less clear relationship and less variation in CWD in general. In all regions, α_{et} was close to 1 with large water input ($w_{\text{in}} > 5$ mm). In arid and semiarid climates, α_{et} takes a larger range of values, decreasing with CWD (dryer soil). The 10–90$^{\text{th}}$ percentile spread

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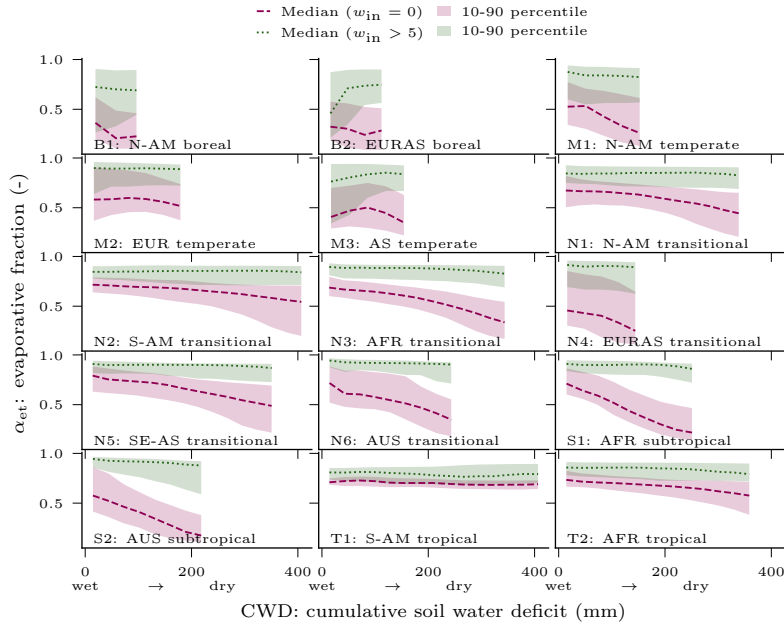
Figure 7. Relationship between evaporative fraction (α_{et}) and cumulative soil water deficit (CWD) for different hydroclimatic regions. The lines shows the respective percentiles of the spatio-temporal conditional distribution $P(\alpha_{et} \mid C \in B_i)$, where C is the cumulative soil water deficit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

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

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

3.3 Terrestrial water storage composition

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

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

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

[*]GPCP for H2M, else WFDEI.

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

510   Not only the spatial patterns of the TWS partitioning show large variations. Figure 9 illustrates the differences in amplitude and timing for the global time series and their decomposition into MSC and IAV. For the seasonal TWS signal, the amplitudes are qualitatively similar, and the main contribution comes from the snow. H2M, SURFEX-TRIP, and PCR-GLOBWB show a soil moisture slightly delayed to the snow seasonality, and the groundwater peak setting in in the late northern spring. W3RA shows very similar soil moisture and groundwater curves, being slightly delayed to the snow seasonality, and LISFLOOD
515   simulates groundwater and soil moisture in alternating cycles with only little variability. The IAV timings of the components are more consistent, but the amplitudes differ significantly across the models. The H2M attributes most TWS$_{IAV}$ to variations in soil moisture, while groundwater dominates the signal for PCR-GLOBWB. Note that the groundwater component also includes the surface water storage for the latter. Also, SURFEX-TRIP and PCR-GLOBWB both show a large global negative IAV anomaly from 2005 to 2006 and a positive one from 2008 to 2010, which are not observed by GRACE.
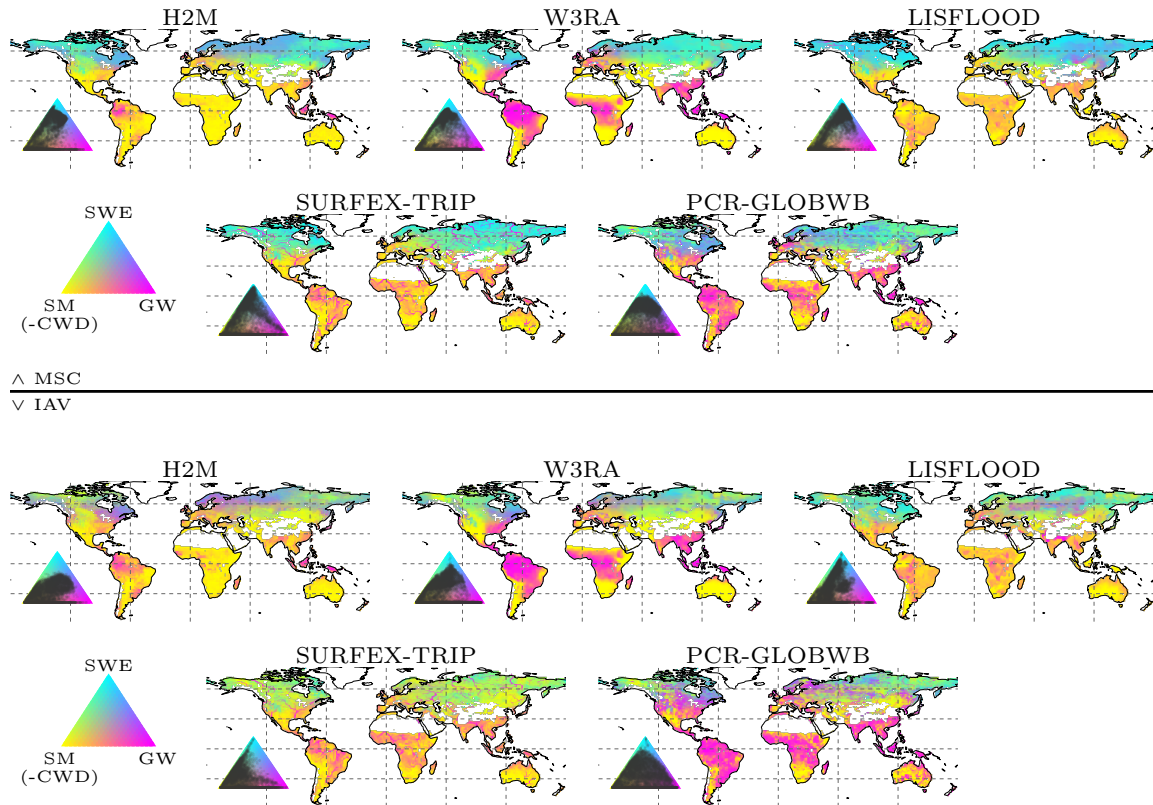
Figure 8. Terrestrial water storage (TWS) variation partitioning into soil moisture (SM, corresponding to negative modeled cumulative water deficit, CWD), groundwater (GW), and snow water equivalent (SWE) variation based on the validation period for the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs). The top panels show the partitioning of the mean seasonal cycle (MSC), the bottom the interannual variability (IAV). The map colors correspond to the mixture of the contributions of the three 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. The decomposition is done based on the years 2003 to 2012.

4 Discussion

In this section, we briefly discuss the model performance and then assess the plausibility of a set of hydrological responses in H2M. We discuss the machine-learned relationship between CWD and runoff generating processes, followed by an analysis of the CWD-α_{et} (evaporative fraction) relationship. Then, we shed some light into the contrast of TWS composition between H2M and GHM simulations. Finally, we discuss general challenges and opportunities of the hybrid approach.
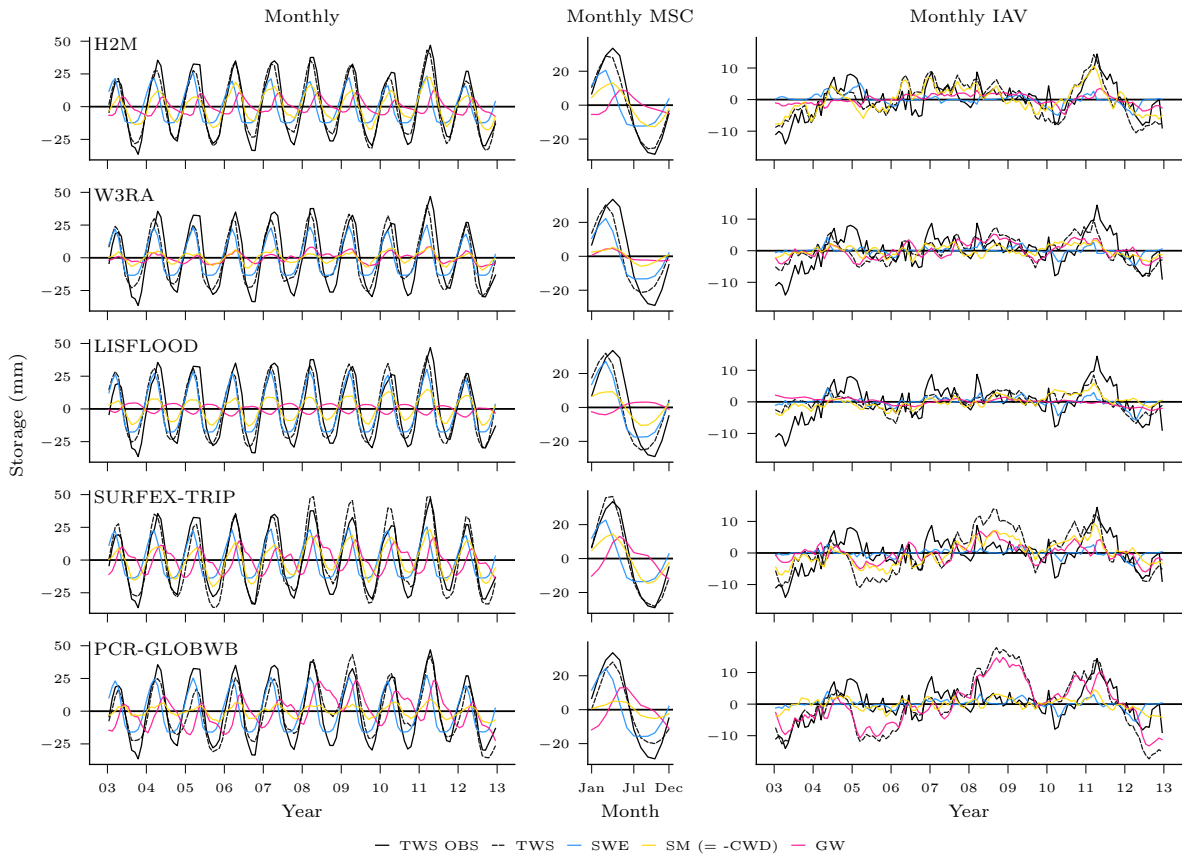
Figure 9. Global variability of the terrestrial water storage (TWS) and the components snow water equivalent (SWE), soil moisture (SM), and groundwater (GW) for the hybrid hydrological model (H2M) and the process-based global hydrological models (rows). Note that SM corresponds to negative modeled cumulative water deficit (CWD) in H2M. For reference, the TWS observations are shown (TWS OBS). The monthly signal (left) and its decomposition into the mean seasonal cycle (MSC, center) and the interannual variability (IAV, right) are shown (columns). The time series represent the global signal, i.e., the data were aggregated using the cell size weighted average per time step, only cell time steps present in all model simulations were used. The y-scale is consistent in columns but varies across the signal components. The training and test period is shown for the complete years 2003 to 2012. Note that surface storage is included in the groundwater component for the models SURFEX-TRIP and PCR-GLOBWB.

4.1 Model performance

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

The TWS seasonality was reproduced well by H2M, except for extremely arid climates, with a low signal-to-noise ratio in observation, resulting in poor NSE values but also small RMSE and decent Pearson's correlation. The largest errors occur in humid regions with a stark TWS seasonality and large runoff rates, e.g., the Amazon basin, central Africa, and Southeast Asia (Fig. A1). This may be related to the missing representations of lateral flow or surface water storage variations in general, which can be important TWS contributions in humid environments (Kim et al., 2009; Scanlon et al., 2019).

A near-perfect fit was achieved for the globally averaged SWE seasonality (Fig. 5) while the local performance varied strongly across regions with the poorest performance in extremely cold tundra (Fig. B2). This is likely related to 1) the known saturation of the GlobSnow SWE observations starting at values of 100 to 120 mm (Larue et al., 2017; Luojus et al., 2010). Consequently, we observe an overestimation of SWE in H2M due to an artifact of the reference data set (A2). 2) In addition, the GPCP forcing product is known to overestimate snowfall due to over-correction of snowfall under catch, especially in high-latitude regions with low station density (Behrangi et al., 2016; Panahi and Behrangi, 2019). The sensitivity of SWE_{IAV} to precipitation forcing data is highlighted by substantially better agreement with GlobSnow when H2M was forced with the WFDEI dataset (Fig. D1 in Appendix D).

4.2 Model interpretability

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

4.2.1 Hydrological responses

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

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

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

et al., 2016, 2017; Koirala et al., 2017), and thus the hybrid approach can fill in critical process gaps and long-standing physical modeling challenges.

The learned relationship between evaporative fraction (α_{et}) and soil dryness (Fig. 7) is generally consistent with the "demand-supply" framework for evapotranspiration (Budyko, 1974). Under wet conditions, ET scales with atmospheric demand represented by net radiation, while evaporative fraction declines with increasing dryness which is most clearly seen in the semi-arid regions of Australia and Africa. The learned relationship between α_{et} and soil moisture response functions appear to be rather gradual as opposed to an idealized piece-wise function with a clear soil moisture threshold that is still frequently employed in process models (Seneviratne et al., 2010; Schwingshackl et al., 2017). However, an about constant, potential evaporative fraction was predicted when there was substantial rain (or snowmelt), independent of the soil moisture state (green lines in Fig. 7). This shows that the model implicitly accounts for wetting of the top soil layers, which alleviates water stress even though it represents soil moisture (expressed as negative CWD) as a single bucket. The specific response of evaporative fraction predicted by H2M varies substantially between regions and within regions indicated by the shading in Fig. 7. Vegetation storage capacity has long been identified as a key uncertainty in process models in controlling soil moisture stress responses (Ichii et al., 2009). Our approach in H2M avoids such explicit parameterizations of relatively less understood physical processes, and its effectiveness is supported by better performance of H2M in simulating TWS variations in tropical and subtropical regions compared to GHMs (Sect. 3.1) despite its simple overall structure.

4.2.2 Terrestrial water storage composition

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

The dominant contribution of the SWE to seasonal cycle of TWS in the high latitudes (Fig. 8 & 9), 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 uncertainties in the input precipitation and SWE observations. Despite regional differences, the models also consistently attribute most of the TWS seasonal and interannual variability to soil moisture in arid and semi-arid regions (Fig. 8). The dominance of soil moisture is plausible in these regions, as the potential evapotranspiration is high and precipitation is low and infrequent or strongly seasonal (Nicholson, 2011). Given the absence of secondary moisture sources such as lateral flow and a lack of deep-rooted plants, most of the storage variations occur within a shallow soil depth (Grayson et al., 2006).

In other regions, the partitioning between groundwater and soil moisture variability is less clear. On both the seasonal and interannual scales, groundwater contributions to TWS correlate with humidity at the global scale (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. 9) co-occurring with low evaporative demand due to low temperatures, irradiation, and vegetation productivity, which results in a large water surplus (Jasechko et al., 2014). The boreal regions with stronger soil moisture contribution are the ones affected by permafrost, where most of the vertical movement is limited to the thawed top soil and horizontal baseflow is usually lower than in non-permafrost soils (Bui et al., 2020). Thus, the patterns diagnosed by H2M are plausible. It must be noted, however, that significant drainage of the surplus water happens via river flows and lateral transport, which are not represented in H2M.

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

In the arid-to-wet transition regions of Africa, H2M diagnoses only marginal groundwater variability compared to larger amplitudes in the GHMs. The H2M resolves the water balance mainly using soil moisture variations, i.e., through soil recharge and evapotranspiration, while the soil overflow was negligible. While the patterns found by H2M are within those of GHMs in most regions, the notable strong soil moisture contribution in tropical savanna and humid subtropical climates is 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 texture map and associated vegetation, soil thermal and hydraulic properties. Often, the said parameters are uncertain and may not necessarily represent a process at spatial scale of GHMs (scale mismatch) or within grid or catchment variabilities (sub-grid to local heterogeneity). Thus, simple heuristics have been used to parameterize hydrological processes, which can, in reality, be of high complexity (Beck et al., 2016). It has been suggested that GHMs underestimate the land water storage capacity in general and that especially the variability in deeper soil is too low (Zeng et al., 2008). In addition, the link between deeper soil layers and plant transpiration through root water uptake is often not represented adequately in GHMs (Jackson et al., 2000), although such effects have been found to play an important role in below-surface water variability (e.g., Kleidon and Heimann, 2000; Koirala et al., 2017). Compared to the GHMs, H2M provides a novel avenue on which storage variations are less-bound by presumably ad-hoc 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

630 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 represent all the below-ground water that will be returned to the atmosphere by root water uptake and transpiration at some point. As a possible consequence, H2M

635 diagnoses larger soil moisture in transitional and especially in the subtropical regions, but more evidently, smaller groundwater variability.

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

640 significant impact on the TWS composition (Kim et al., 2009).

4.3 Challenges and opportunities

The data-driven character of the H2M offers a set of opportunities but is accompanied by challenges. The H2M makes use of observational data streams that are not typically used in GHMs. However, to retain interpretability of the predicted coefficients, the model structure must be kept simple: The model flexibility, thus, needs to be compensated with a simple causal model

645 structure. Still, the H2M offers a great opportunity to study the hydrological cycle from a different viewpoint that is strongly footed on the observation-based data sets, that is growing in availability at an unprecedented rate in the era of Earth observation.

An outstanding challenge in hybrid hydrological modeling is the representation of model uncertainties, which would allow for a targeted model development. The hydrological pathways in H2M are rather simple compared to GHMs, but the model still expresses a high data-adaptivity as demonstrated. While GHMs usually represent a wide range of hydrological sub-processes

650 (e.g., infiltration, preferential flow, topographical runoff-runon), the hybrid model integrates them to a few response functions and the model complexity and interactions within is, so to speak, outsourced to the neural network. Still, missing representations of storage components (e.g., surface storage) and hydrological pathways (e.g., streamflows) limit the model flexibility and, to a certain extent, can corrupt the other latent variables as the model tries to accommodate for missing processes. Thus, the estimated coefficients in the current H2M implementation should be treated with some skepticism. At the same time, the

655 relaxation of assumptions can be seen as an opportunity, as the prior knowledge may be wrong or incomplete. The impact of trading prior knowledge and model complexity with more flexibility and data-drivenness on model uncertainties is a key question that needs to be investigated further.

As the model behavior emerges largely from the observational data constraints, the hybrid approach constitutes a novel technique for studying TWS variations. While purely data-driven approaches (see Andrew et al. (2017) for an overview) are

660 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, principles and constraints. GHMs, themselves, largely rely on prior knowledge, which may be false or incomplete, and the model parameterization is usually not resolved regionally, resulting in model uncertainties (Beck et al., 2016) which are eventually

expressed in the disagreement among model simulations. The hybrid model can be seen as a compromise between the purely data-driven and the physically-based approaches, as physical principles (e.g., mass conservation) are respected, but qualitative assumptions on the processes are still used.

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. 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 data constraints or better physical processes in order to avoid parameter equifinality issues that lead to the same or similar model responses across a large range of parameter values. It is well possible that the decomposition into CWD and GW is not properly constrained under some circumstances, e.g., in ecosystems that are not water limited. Here, either the groundwater or the soil moisture may be restored as needed (due to frequent precipitation) to match the observation of terrestrial water storage. More research is needed to address these problems. In particular, a complementary development of application-based models as presented in this study, and smaller-scale, better constrained exercises to advance hybrid modeling can be a viable alternative.

The rapid development of novel products opens interesting opportunities, like a daily TWS product (Kvas et al., 2019) can help to better constrain sub-monthly water processes. Furthermore, the upcoming Surface Water and Ocean Topography (SWOT) mission, which is targeted at observing surface water storage variations (Biancamaria et al., 2016), could be useful to solve current shortcomings of the H2M. In addition, parameters estimated by other approaches, such as the upscaled baseflow index (Beck et al., 2013), offer interesting independent constraints that allow adding 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 representing such uncertainties (Reichle, 2008) and can even be extended to incorporate model parameter estimation (Moradkhani et al., 2005), i.e., learning physical processes as in the hybrid approach presented here. In contrast to data assimilation that often target improving prediction skills, the goal of hybrid modeling is to develop a generalizable model, which can be applied beyond the specific forecasting task in data assimilation. Nevertheless, non-parametric machine learning approaches can also be included into data assimilation as discussed in Geer (2021).

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 regions. Such endeavors would also

allow for bringing in established global datasets of river discharge measurements such as provided by the Global Runoff Data Centre (GRDC, Fekete et al., 1999).

5 Conclusions

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

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

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

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

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

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

Code and data availability. The H2M and its training are implemented in *PyTorch 1.5* (Paszke et al., 2017), an open-source deep learning framework for the *Python* programming language. 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 data used as model input, but all datasets are referenced in the manuscript.
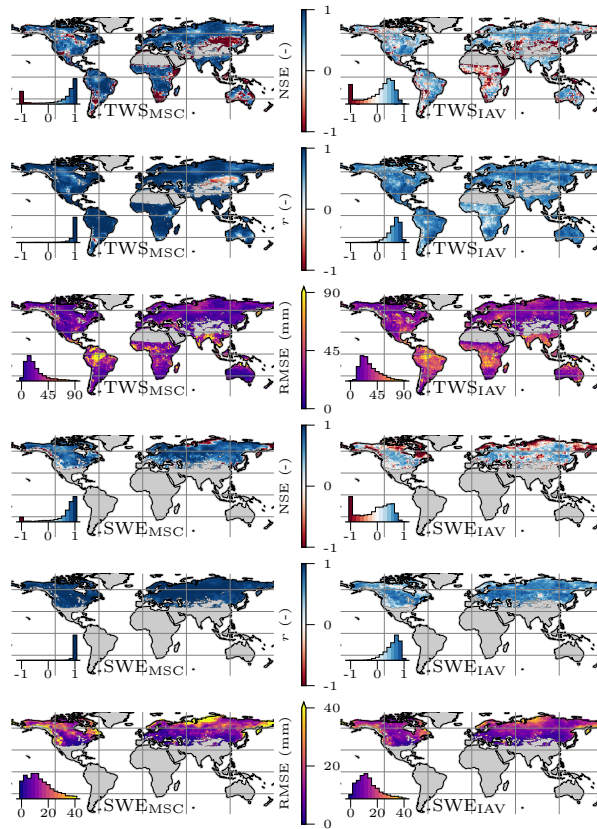
Figure A1. Local model performance for terrestrial water storage (TWS) and snow water equivalent (SWE) on the mean seasonal cycle (MSC) and the interannual variability (IAV) within the test period (2009 to 2014). The Nash–Sutcliffe model efficiency (NSE), Pearson correlation (r) and Root Mean Square Error (RMSE) are shown. The inset plots show the cell area–weighted histogram of the map values.

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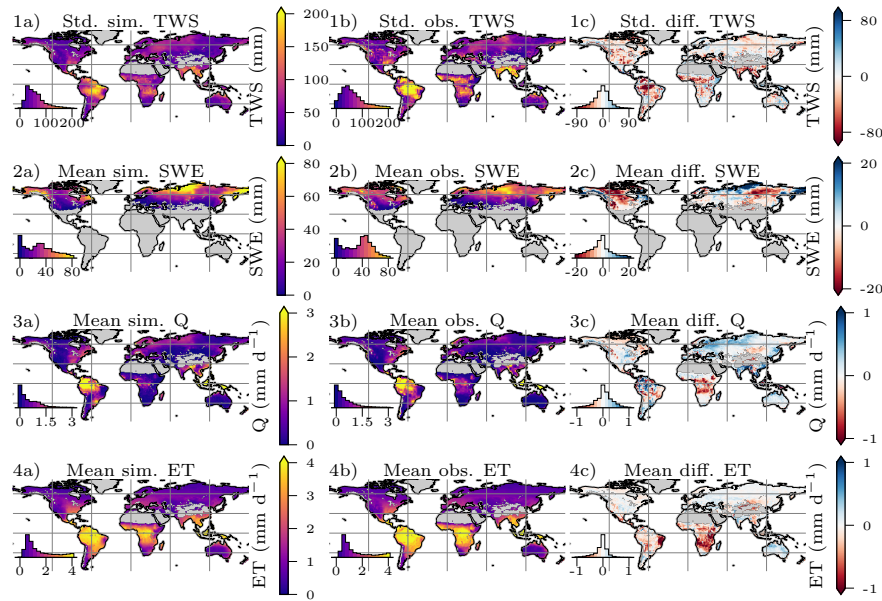
Figure A2. Mean a) simulated, b) observed, and c) difference of simulated minus observed (positive means simulated is larger) terrestrial water storage (TWS, 1a–c), snow water equivalent (SWE, 2a–c), total runoff (Q, 3a–c), and evapotranspiration (ET, 4a–c). Note that for the TWS, the standard deviation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

735 of large rivers (e.g., Amazon, Ganges, Mississippi, Niger, or Yenisei) and may be caused by missing representation of river routing. The SWE is overestimated in the extremely cold regions of North America and Northeast Asia, and underestimated in Tundra regions. Average Q is largely underestimated in Central Africa, and slightly overestimated in northwestern Eurasia, central Amazon, and coastal regions of Australia and East Asia. ET, finally, is underestimated by the model, prominently in most of Subsaharan Africa and East Brazil, while no major biases are present in other regions.

740 **Appendix B: Regional comparison of simulated time series**

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

Most models manage to reproduce the SWE_{MSC} well with an $e_{\text{NSE}} > 0.5$, while the SWE_{IAV} performance is more variant and lower in general (Fig. B2). We note a phase difference between the model simulations and observations that is most notable in
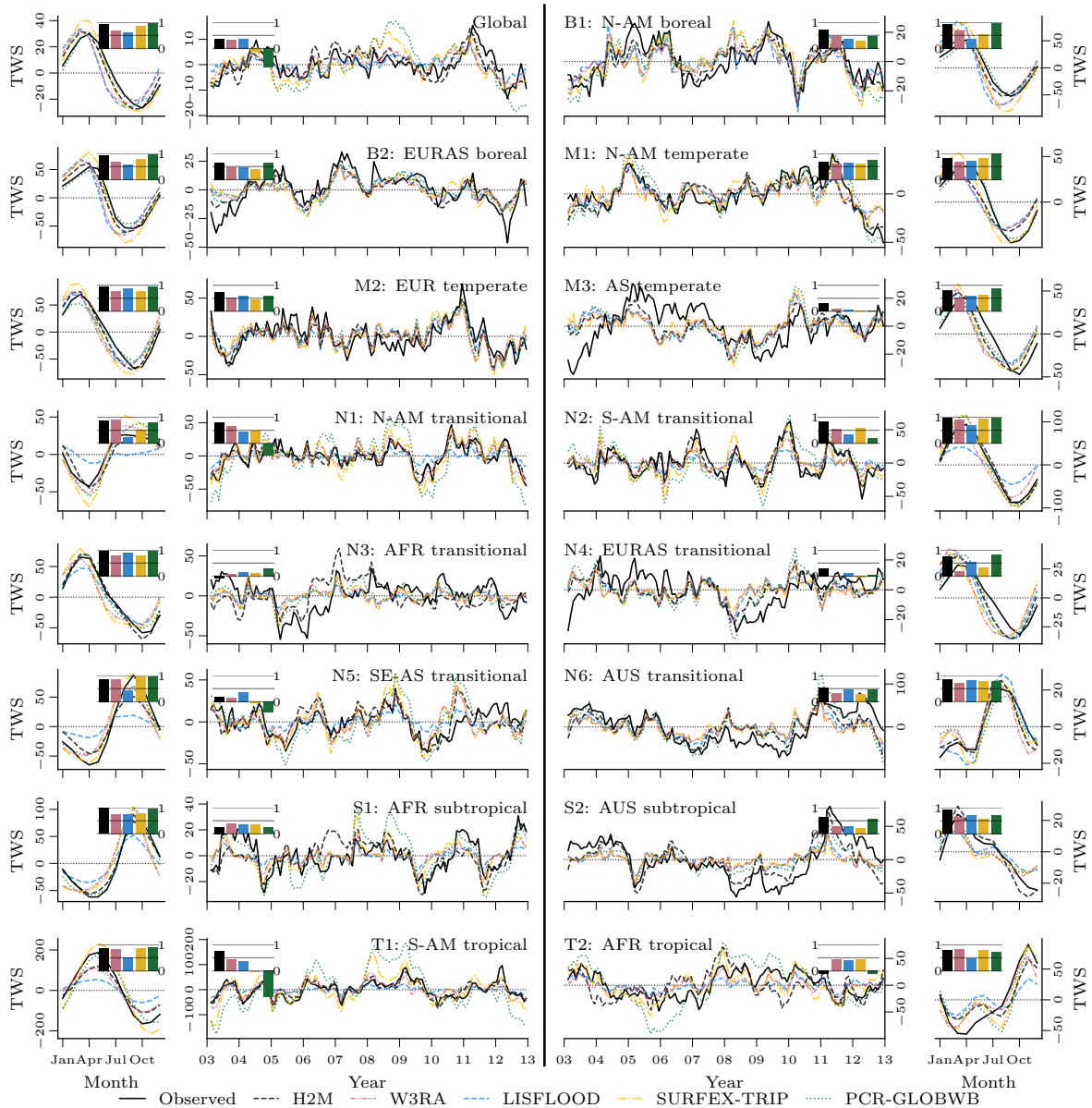
Figure B1. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the terrestrial water storage mean seasonal cycle (TWS$_{MSC}$, outer columns) and interannual variability (TWS$_{IAV}$, center columns) in mm for hydro-climatic regions (Fig. 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.
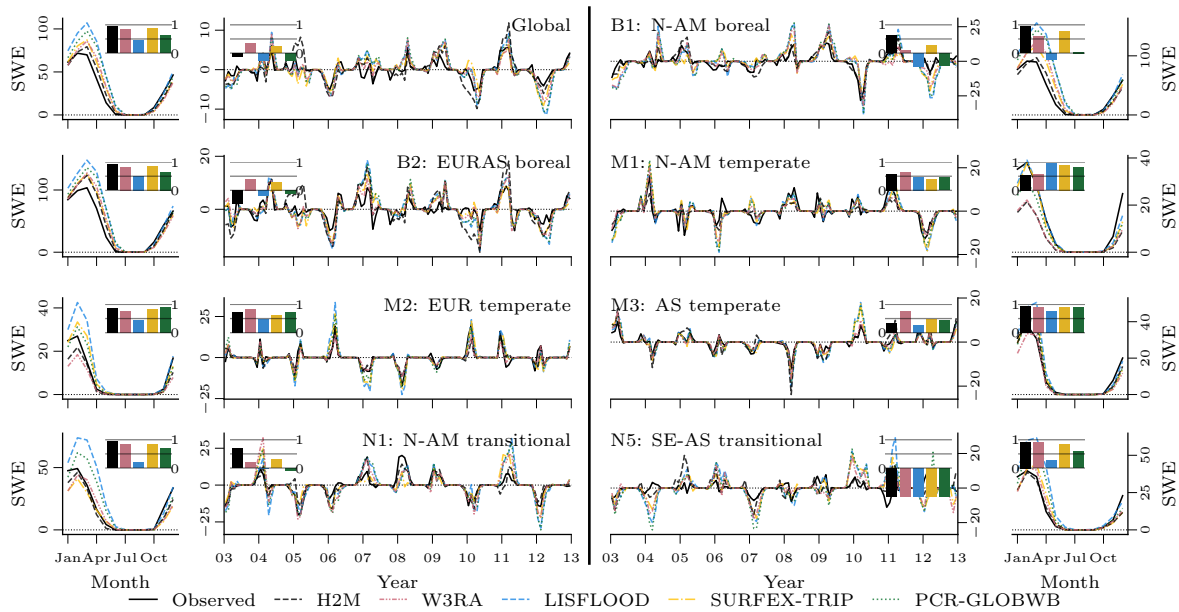
Figure B2. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the snow water equivalent mean seasonal cycle (SWE_{MSC}, outer columns) and interannual variability (SWE_{IAV}, center columns) in mm for hydro-climatic regions (Fig. 2). The 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.

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
750   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.

The regional scale seasonal anomalies of simulated soil moisture (corresponding to negative CWD in H2M) and GW show a more detailed picture of the model variabilities (Fig. B3). The global scale SM amplitude of H2M is larger than the one of the
755   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.
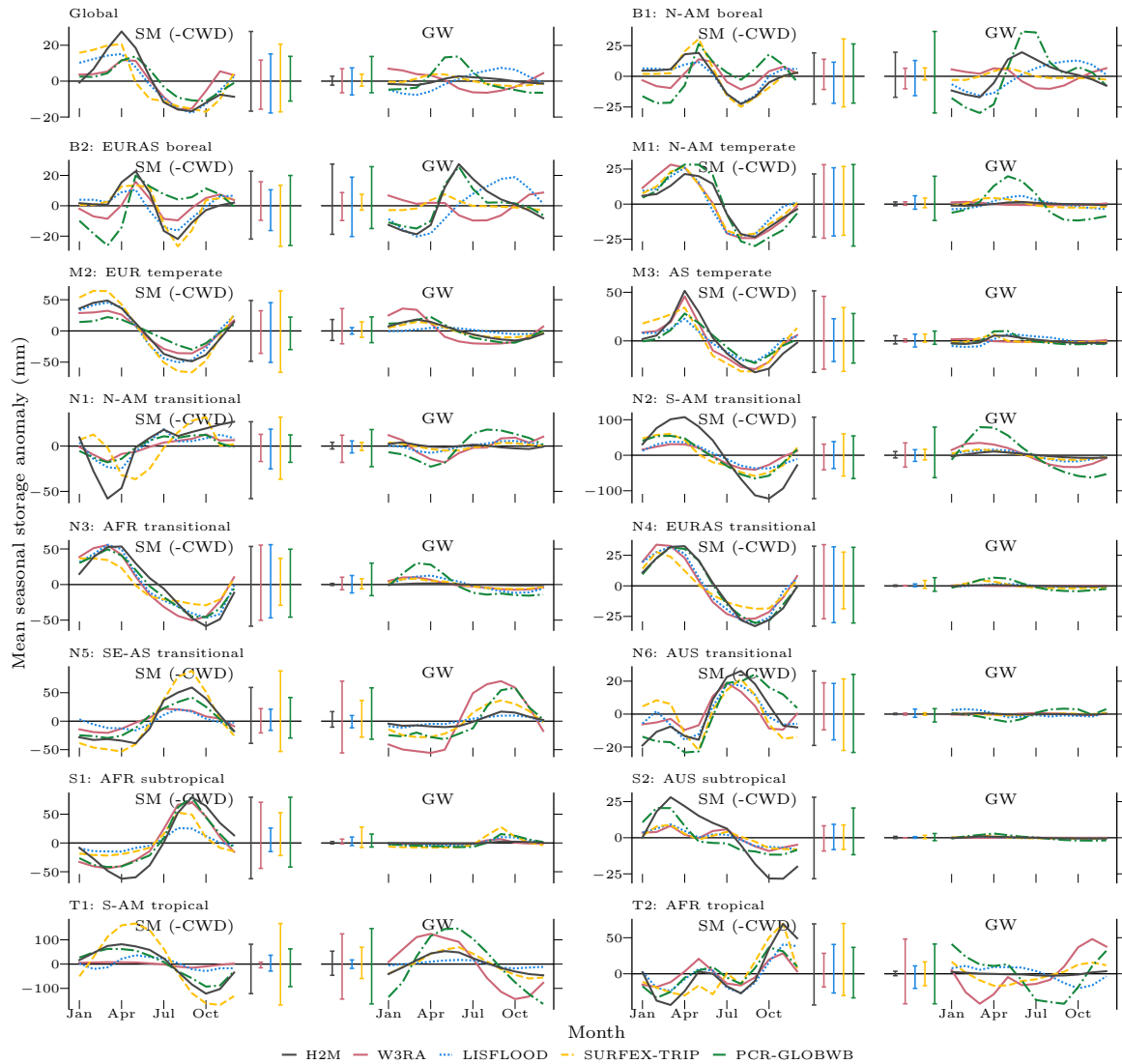
Figure B3. Global and regional mean seasonal anomalies of soil moisture (SM) and groundwater (GW) for the hybrid model (H2M) and the process-based global hydrological models. Note that SM corresponds to negative modeled cumulative water deficit (CWD). Ranges from the minimum to the maximum value per model are shown next to the seasonal cycle as vertical lines. The regions are shown in 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 time steps from 2009 to 2014. Note that the y-scale is consistent within, but differs across regions.
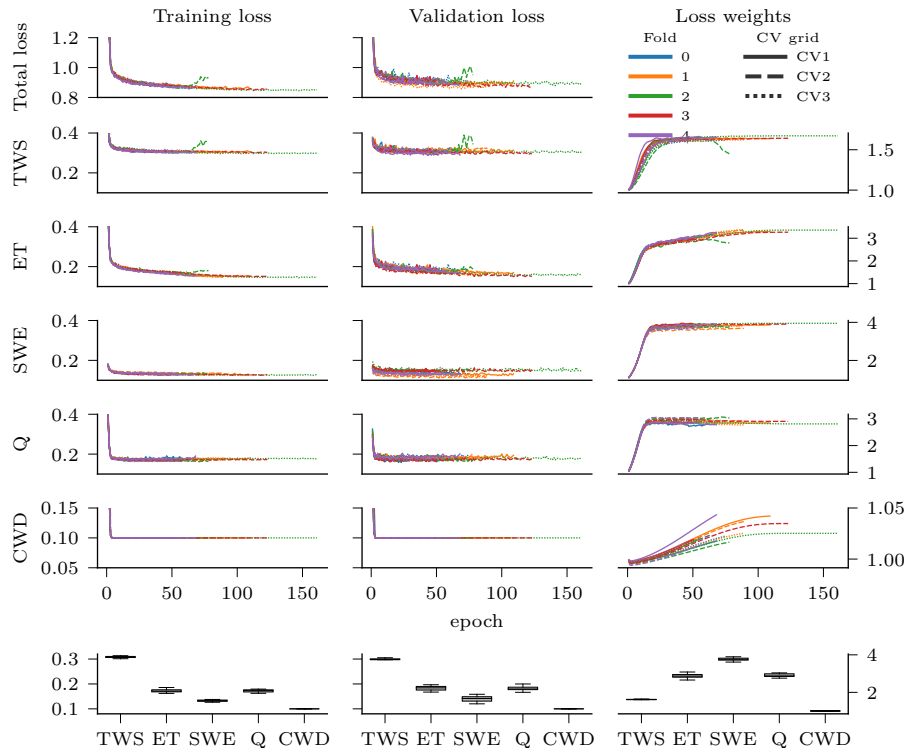
Figure C1. Model training process for the cross-validation runs. The left and central columns represent the unweighted total and variable-specific MSE loss. The right column shows how the task weights developed over training time. The x-axis represents the number of iterations through the training set ("epochs"). The bottom row contains the column-wise distribution of the variables losses (or weights) at the end of the model optimization. Note that for the soft constraint on CWD, a bias of 0.1 was added, i.e., 0.1 is the optimum.

Appendix C: Model optimization

The model optimization within the cross-validation setting is shown in Fig. C1. The learning process was stable in most cases and a smooth model convergence was achieved. Only one run (fold 2, CV2) was unstable as the training collapsed. Due to the early stopping mechanism, however, the model from the best validation loss is restored and used for the test set prediction. The loss and weight ($w = \frac{1}{2\sigma^2}$, where σ is the task uncertainty, see Sect. 2.3.3) distributions at optimum across cross-validation runs were stable (bottom row of boxplots in Fig. C1). The generalization loss from the training to the validation loss is minimal, although a slightly larger spread of the validation losses can be observed. The 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 were stable across cross-validation runs. The weights are difficult to interpret, as they do not directly translate to inverse variable uncertainty (Kendall et al., 2018) but also depend on the variable variance (although the loss is calculated on standardized data). From the boxplots in Fig. C1, we can see that variables with a lower loss are given more
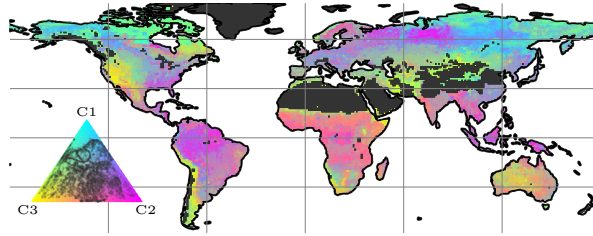
Figure C2. The t-distributed stochastic neighbor (t-SNE) reduction to three dimensions (C1-3) of static variable encoding (originally 12 dimensions, ρ_{enc} in Fig. 1) of one cross-validation run. The encoding is a low-level representation of the static inputs, i.e., soil and land-cover properties, learned by a neural network. The inset ternary plots show the distribution of the map values.

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.

775   Part of the model tuning involved optimization of the sub-network FCNN[1] (Fig. 1), extracting features from the static variables which are then fed into the recurrent neural network. We visualized the outputs (ρ_{enc} in Fig. 1) of the FCNN[1] 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. C2) reveals patterns that seem very familiar: the components align with patterns of biomass, vegetation type and aridity. Note that the t-SNE algorithm is non-deterministic

780 and can yield vastly different results depending on chosen hyper-parameters. Also, the reduction to three dimensions only reveals the major gradients and does not represent the entire variability.

Appendix D: Model forcing with WFDEI

To test the impact of the forcing datasets, the model was trained on the WFDEI forcings (Weedon et al., 2014) as used in the eartH2O ensemble. The performance (Fig. D1) in respect to TWS was almost identical with slightly larger NSE on the global

785 signal and lower NSE on local level when using WFDEI. The NSE of SWE was larger with WFDEI, especially for the IAV. Due to the similar performance, we conclude that the impact of the forcings is negligible and the results are robust in regards to them.
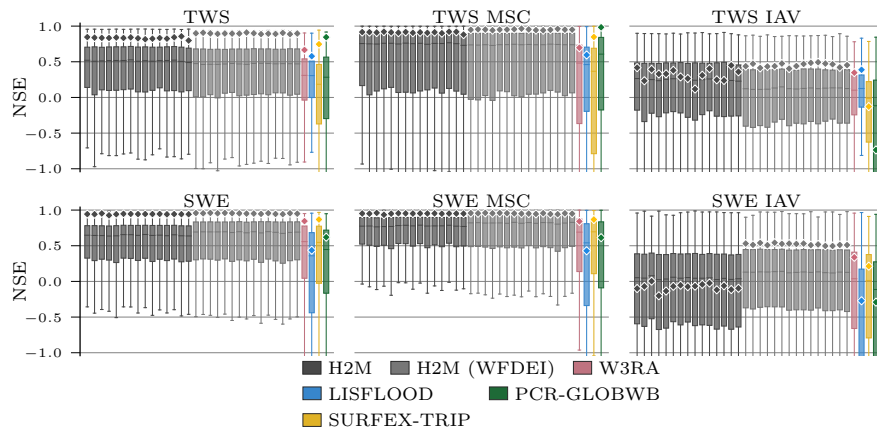
Figure D1. Global and local grid cell–level Nash–Sutcliffe model efficiency coefficient (NSE) of the hybrid hydrological model (H2M) and the process-based global hydrological models (GHMs) for the terrestrial water storage (TWS) on top and the snow water equivalent (SWE) at the bottom. The gray bars represent the cross-validation runs using the forcings described in Section 2.1.1 (dark grey, "H2M"), and using the WFDEI forcings as used for in the eartH2O ensemble (light grey, "H2M (WFDEI)"). The ⋄-markers show the global (spatially averaged signal) model performance, the boxes represent the spatial variability of the local cell-level performance. The y-axis was cut at -1 due to some large negative NSE values. The panels show the model performance 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 2009 to 2012 on monthly time scale. Note that deviations from the numbers reported in Tab. 3 are due to different time ranges.

Appendix E: Model pseudo-code

Algorithm 1 The training loop of the hybrid hydrological model.

1: $\phi, \beta, \sigma \leftarrow initialize()$ # Initialize model weights ϕ, global constants β, task uncertainties σ

2: **while** not converged **do**

3:    $cells \leftarrow sample_{\text{cells}}(gridcells, n)$ # sample n gridcells

4:    $m_{\text{sim}} \leftarrow meteo[cells]$ # select cells from forcings

5:    $m_{\text{spinup}} \leftarrow sample_{\text{spinup}}(m_{\text{sim}}, 5)$ # sample 5 random years

6:    $m \leftarrow concat(m_{\text{spinup}}, m_{\text{sim}})$ # concatenate

7:    $\rho \leftarrow static[cells]$ # select cells from static

8:    $y \leftarrow target[cells]$ # select cells from targets

9:    $c, h \leftarrow zeros(100)$ # initialize LSTM hidden states

10:    $s \leftarrow zeros(3)$ # initialize physical storages

11:    $loss \leftarrow 0.0$ # initialize loss

12:    $\rho_{\text{enc}} \leftarrow FCNN^1(\rho)$ # compress static encodings

13:    **for** $t \in \{1, \ldots, T\}$ **do**

14:      $c, h \leftarrow LSTM(c, h, s, m[t], \rho_{\text{enc}})$ # update LSTM states

15:      $\alpha \leftarrow FCNN^2(h)$ // get coefficients

16:      $s, f \leftarrow hydro(s, m[t], \alpha, \beta)$ # run phys. model, get storages s and fluxes f

17:      $\hat{y} \leftarrow collect(s, f)$ # collect target variables

18:      **if** $t \notin spinup$ **then**

19:        $loss \leftarrow loss + MSE(\hat{y}, y[t], \sigma)$ # add weighted loss to previous loss

20:      **end if**

21:    **end for**

22:    $\phi, \beta, \sigma \leftarrow update(\phi, \beta, \sigma, loss)$ # update parameters

23: **end while**

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.

790

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 pre-processing and the colleagues from the MPI for Biogeochemistry and the TU Munich for the inspiring discussions.

795 **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
800   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
805   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,
810   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.

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.

815 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,
820   https://doi.org/10.1016/j.isprsjprs.2014.09.002, 2015.

de Bézenac, E., Pajot, A., and Gallinari, P.: Deep learning for physical processes: Incorporating prior scientific knowledge, Journal of Statistical Mechanics: Theory and Experiment, 2019, 124 009, https://doi.org/10.1088/1742-5468/ab3195, 2019.

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

825 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.

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

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

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

Falkner, S., Klein, A., and Hutter, F.: BOHB: Robust and efficient hyperparameter optimization at scale, arXiv preprint, https://arxiv.org/abs/1807.01774, 2018.

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.

Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: Global, composite runoff fields based on observed river discharge and simulated water balances, https://csdms.colorado.edu/wiki/Data:GRDC, 1999.

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.

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.

870   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
875   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.

880   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
885   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., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlström, A., Arneth, A., Camps-Valls, G., Ciais, P., Friedling-
          stein, P., et al.: Compensatory water effects link yearly global land CO_2 sink changes to temperature, Nature, 541, 516–520,
          https://doi.org/10.1038/nature20780, 2017.

890   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.

        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,
895   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 simula-
          tions: 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., Tramon-
900   tana, G., et al.: Global distribution of groundwater-vegetation spatial covariation, Geophysical Research Letters, 44, 4134–4142,
          https://doi.org/10.1002/2017GL072885, 2017.

        Körner, M. and Rußwurm, M.: Recurrent Neural Networks and the Temporal Component, Deep Learning for the Earth Sciences: A Com-
          prehensive Approach to Remote Sensing, Climate Science, and Geosciences, pp. 105–119, https://doi.org/10.1002/9781119646181.ch8,
          2021.

905   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.

Kraft, B., Besnard, S., and Koirala, S.: Emulating Ecological Memory with Recurrent Neural Networks, Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, pp. 269–281, https://doi.org/10.1002/9781119646181.ch18, 2021.

Kumar, A.: Stage-Discharge Relationship, in: Encyclopedia of Snow, Ice and Glaciers. Encyclopedia of Earth Sciences Series, edited by Singh, V., Singh, P., and Haritashya, U., Springer, Dordrecht, https://doi.org/10.1007/978-90-481-2642-2_537, 2011.

Kvas, A., Behzadpour, S., Ellmer, M., Klinger, B., Strasser, S., Zehentner, N., and Mayer-Gürr, T.: ITSG-Grace2018: Overview and evaluation of a new GRACE-only gravity field time series, 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, arXiv preprint, https://arxiv.org/abs/1807.05118, 2018.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, arXiv preprint, https://arxiv.org/abs/1711.05101v3, 2017.

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.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., et al.: Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, IPCC, 2014.

Panahi, M. and Behrangi, A.: Comparative analysis of snowfall accumulation and gauge undercatch correction factors from diverse data sets: In situ, satellite, and reanalysis, Asia-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.

945   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.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

950   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.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller,
955   W., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Rodell, M., Famiglietti, J., Wiese, D., Reager, J., Beaudoing, H., Landerer, F. W., and Lo, M.-H.: Emerging trends in global freshwater availability, Nature, 557, 651–659, https://doi.org/10.1038/s41586-018-0123-1, 2018.

Scanlon, B., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoing, H., Lo, M., Müller-Schmied, H., Döll, P., et al.: Tracking
960   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.

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.

965   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.

970   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.

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.

975   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.

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.

980   Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

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

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 vari- able mass distribution with GRACE using spherical cap mascons, Journal of Geophysical Research: Solid Earth, 120, 2648–2671, https://doi.org/10.1002/2014JB011547, 2015.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 50, 7505–7514, https://doi.org/10.1002/2014WR015638, 2014.

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee III, R. B., Smith, G. L., and Cooper, J. E.: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment, 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.