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

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

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

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

ABSTRACT:

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

1. INTRODUCTION

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

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

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

2. GLOBAL DATASETS

2.1 Total Water Storage Anomalies (TWS)

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

2.2 Evapotranspiration (ET)

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

* Corresponding author

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

2.3 Total Runoff (Q)

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

2.4 Snow Water Equivalent (SWE)

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

2.5 Meteorological Forcing

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

2.6 Static Datasets

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

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

2.7 Masking & Bioclimatic Regions

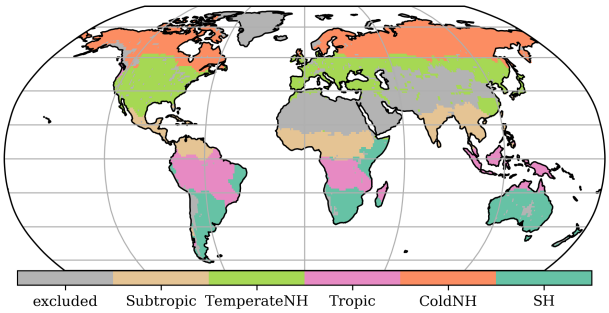


Figure 1. The masked out cells ('excluded') and the bioclimatic regions used for model evaluation: Cold Northern Hemisphere ('ColdNH'), Temperate Northern Hemisphere ('TemperateNH'), 'Tropic', 'Subtropic' and remaining Southern Hemisphere regions ('SH').

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

3. GLOBAL HYBRID HYDROLOGICAL MODELING

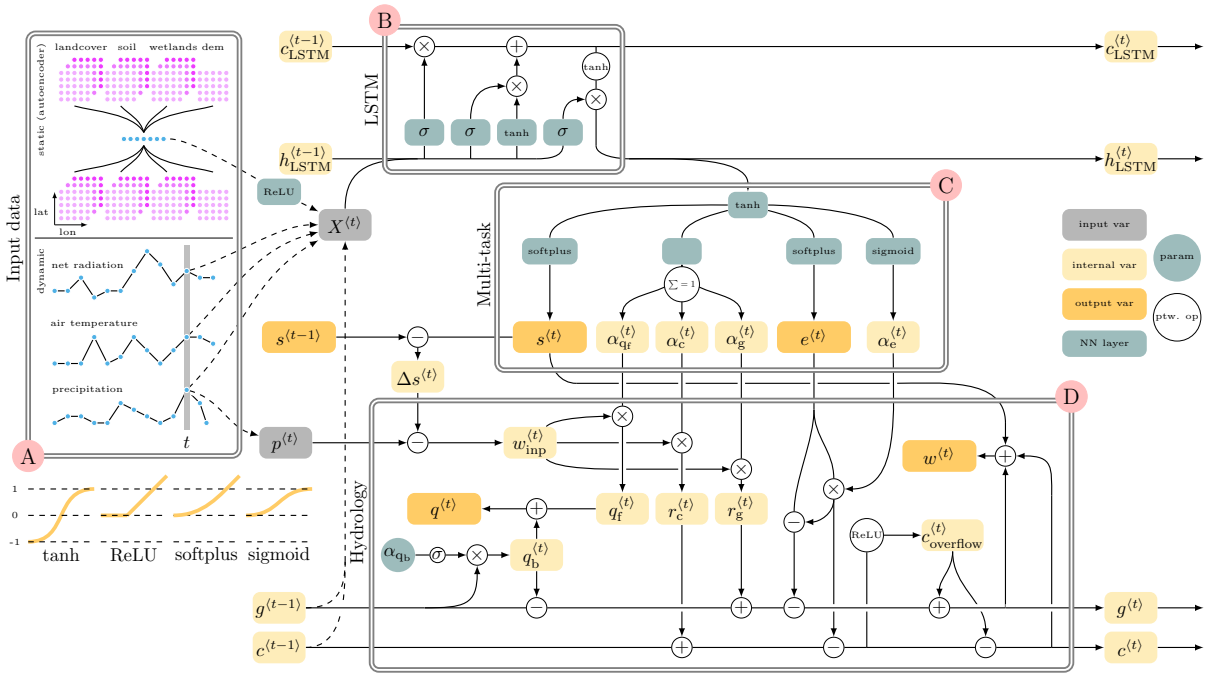
3.1 The Hybrid Hydrological Model

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

3.2 Self-Paced Multi-Task Learing

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

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



A Input data

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

B The LSTM layer

The LSTM updates the hidden states $h_{\text{LSTM}}^{\langle t \rangle}$ and $c_{\text{LSTM}}^{\langle t \rangle}$ at each time-step.

$$h_{\text{LSTM}}^{\langle t \rangle}, c_{\text{LSTM}}^{\langle t \rangle} = \text{LSTM}(h_{\text{LSTM}}^{\langle t-1 \rangle}, c_{\text{LSTM}}^{\langle t-1 \rangle}, X^{\langle t \rangle})$$

C Multi-task layer

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

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

D Water balance model

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

$$\begin{aligned}
c^{\langle t \rangle} &= c^{\langle t-1 \rangle} + \overbrace{\alpha_\text{c}^{\langle t \rangle}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{r_\text{c}^{\langle t \rangle}} - e^{\langle t \rangle}\alpha_\text{e}^{\langle t \rangle} \\
c^{\langle t \rangle} &= c^{\langle t \rangle} - \overbrace{\max(c^{\langle t \rangle}, 0)}^{c_{\text{overflow}}^{\langle t \rangle}} \\
q^{\langle t \rangle} &= \overbrace{\alpha_{\text{q}_\text{f}}^{\langle t \rangle}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{q_\text{f}^{\langle t \rangle}} + \overbrace{g^{\langle t-1 \rangle}\,\text{sigmoid}(\alpha_{\text{q}_\text{b}}) \cdot 0.01}^{q_\text{b}^{\langle t \rangle}} \\
g^{\langle t \rangle} &= g^{\langle t-1 \rangle} - q_\text{b}^{\langle t \rangle} + \overbrace{\alpha_\text{g}^{\langle t \rangle}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{r_\text{g}^{\langle t \rangle}} + \\
&\quad c_{\text{overflow}}^{\langle t \rangle} - e^{\langle t \rangle}(1 - \alpha_\text{e}^{\langle t \rangle}) \\
w^{\langle t \rangle} &= s^{\langle t \rangle} + g^{\langle t \rangle} + c^{\langle t \rangle}
\end{aligned}$$

[a] (acronym, math. symbol)

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

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

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

3.3 Model Selection & Training

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

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

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

3.4 Model Evaluation

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

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

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

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

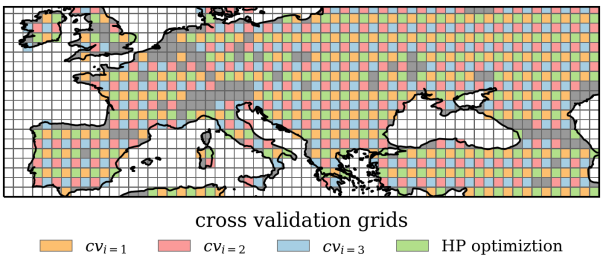


Figure 2. Regional example of the data splitting for the hyperparameter tuning and cross-validation. The grid-cells are split into four mutually exclusive, regular grids (colored). The grid-cells of each set are separated by a buffer to reduce the spatial autocorrelation between the samples. The samples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

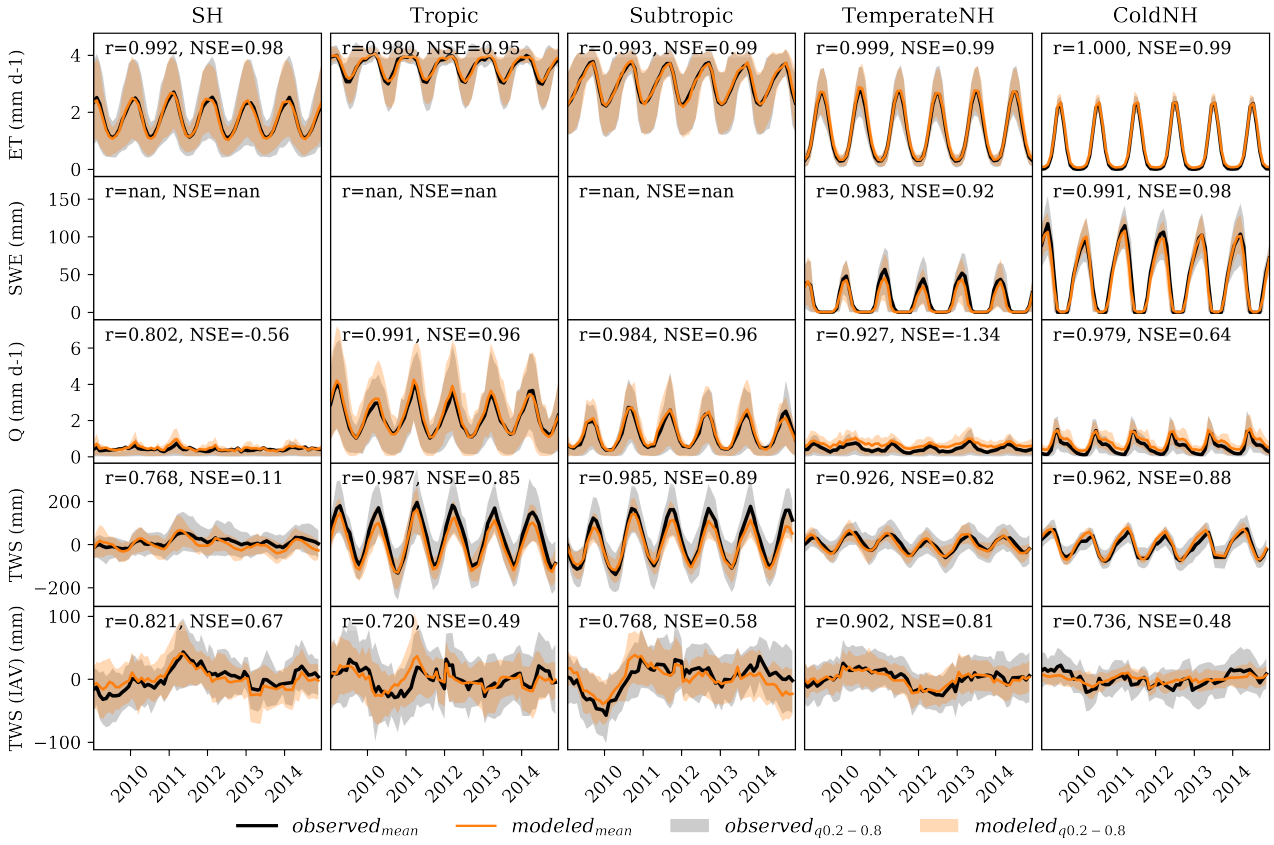


Figure 3. The model performance based on the test set by bioclimatic regions. The four target variables evapotranspiration (ET), snow water equivalent (SWE), runoff (Q), total water storage (TWS), as well as the TWS interannual variability (IAV) are shown. The TWS IAV is calculated as the deviation from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

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

4. RESULTS & DISCUSSION

4.1 Model Performance by Bioclimatic Regions

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

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

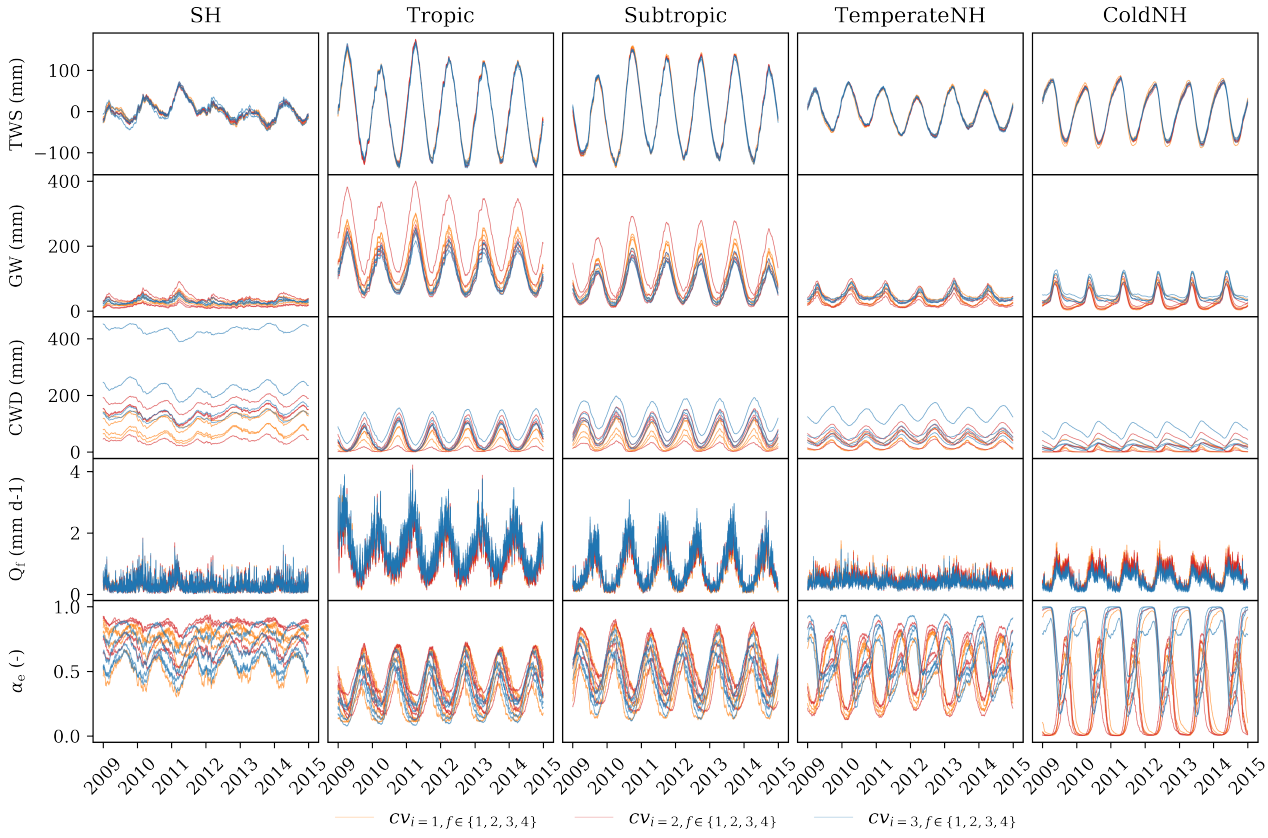


Figure 4. Regional mean time series of repeated model simulations of total water storage (TWS) and the latent variables groundwater (GW), soil cumulative water deficit (CWD), fast runoff (Q_f), and ET partitioning fraction (α_e), defining to what share evapotranspiration is extract form the soil versus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

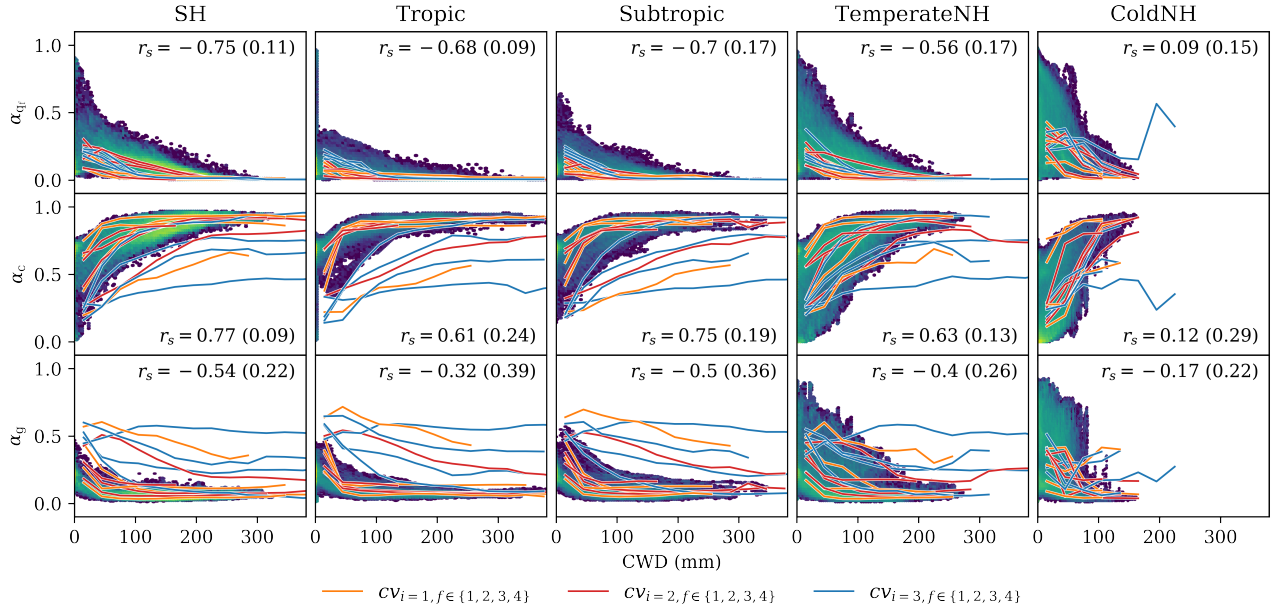


Figure 5. Density plot of the soil cumulative water deficit (CWD) versus the liquid water input (w_{inp}) partitioning fast runoff fraction (α_{q_f}), soil recharge fraction (α_c), and groundwater recharge fraction (α_g). The fractions define how much of w_{inp} goes into the respective fluxes. The relationships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

4.2 Model Robustness & Latent Variables

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

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

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

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

4.3 Limitations

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

5. CONCLUSION

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

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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