Towards global hybrid hydrological modeling by fusing deep learning and a conceptual model

Basil Kraft[1, 2], Martin Jung[1],
Marco Körner[2] and Markus Reichstein[1]
bkraft@bgc-jena.mpg.de

[1]Dept. of Biogeochemical Integration, MPI for Biogeochemistry, Jena, Germany
[2]Dept. of Aerospace and Geodesy, Technical University of Munich, Munich, Germany

EGU, May 2020

Max Planck Institute for Biogeochemistry

Technische Universität München

Introduction

Hybrid modeling

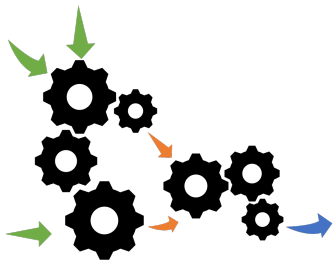
Hybrid modeling

...combines the strengths of physical modelling (theoretical foundations, interpretable compartments) and machine learning (data-adaptiveness)[a]

[a] Reichstein et al. (2019)

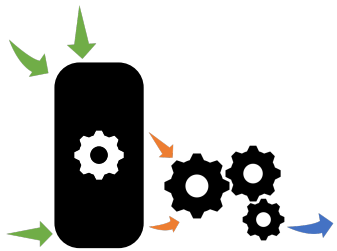
Hybrid modeling

Conceptual

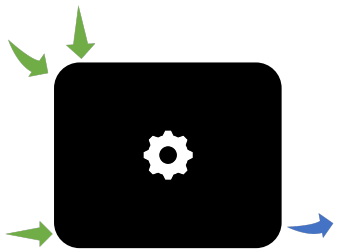


- knowledge driven
- physically interpretable
- extrapolation possible

Hybrid



Black box



- data driven
- less prior knowledge

Table: Hybrid modeling: the combination of knowledge and data driven models.

Example: SST prediction by de Bézenac et al. (2019)

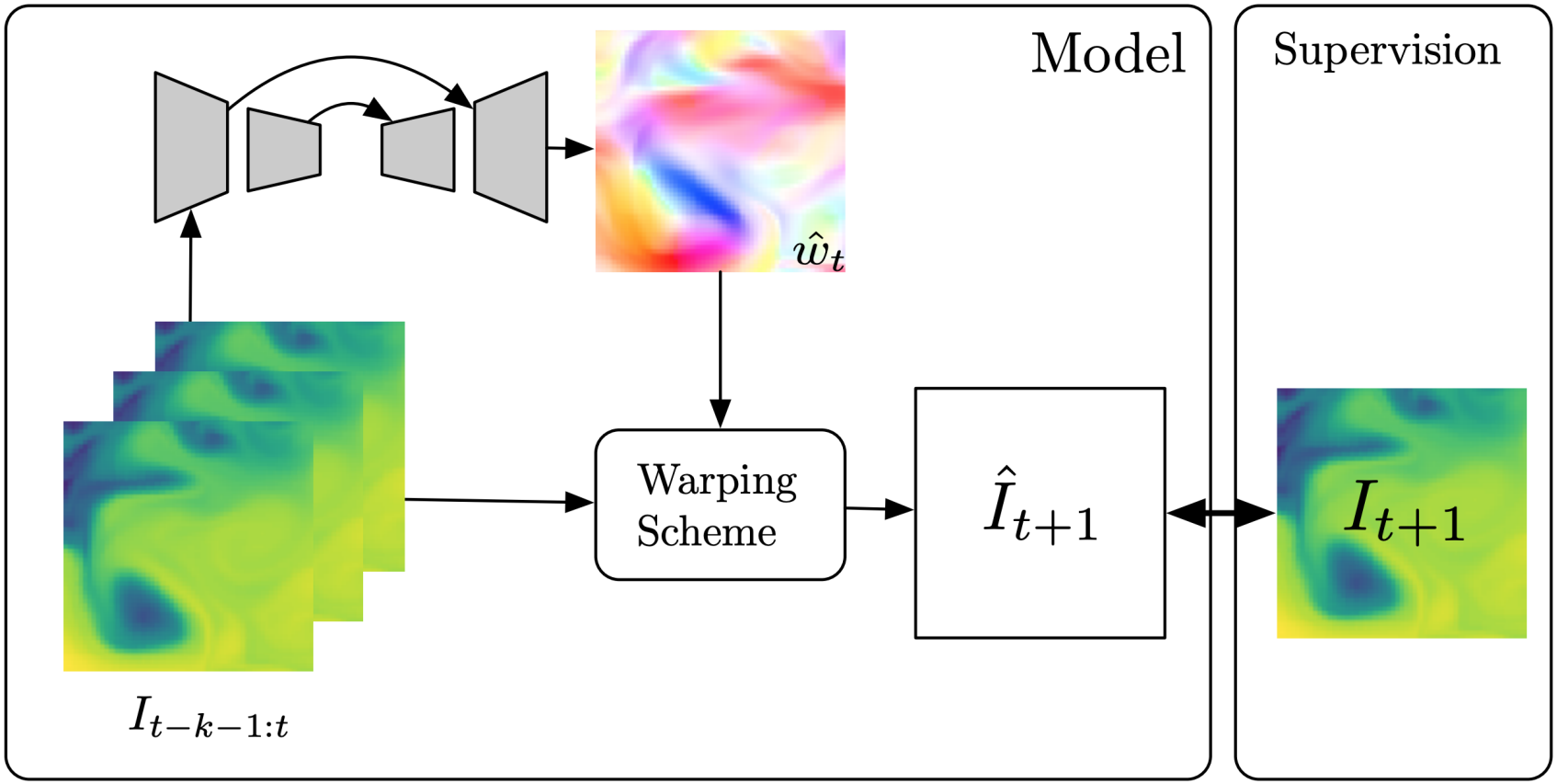


Figure: de Bézenac et al. (2019) used a hybrid model to predict sea surface temperatures (SST). A convolutional encoder-decoder yields a hidden variable, the motion field $\hat{\omega}_t$. $\hat{\omega}_t$ is then fed into a physical model of advection and diffusion, yielding future states of SST. ©IOP Publishing.

Example: SST prediction by de Bézenac et al. (2019)

Advantages of the hybrid approach

- Improved performance by adding prior knowledge (reducing solution space)
- Constraints can be put latent variables (here: $\hat{\omega}_t$)
- Interpretable latent variable (of minor importance in this example)

Motivation

Can we adapt this concept to model large-scale environmental systems, e.g. the global water cycle?

Conceptual water balance models

- based on expert knowledge
- tunable parameters but not *truly* data adaptive
- yet we have large amounts of data

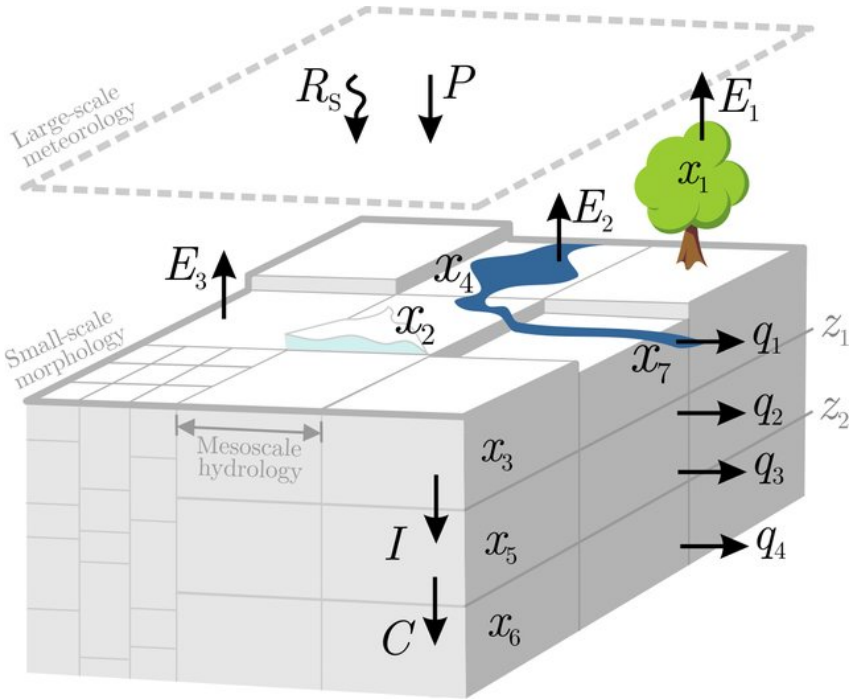


Figure: Mesoscale hydrologic model (mHM), Samaniego et al. (2010).

Deep learning

- data adaptive
- requires low prior knowledge
- specialized models, e.g. long short-term memory (LSTM) network for sequential data

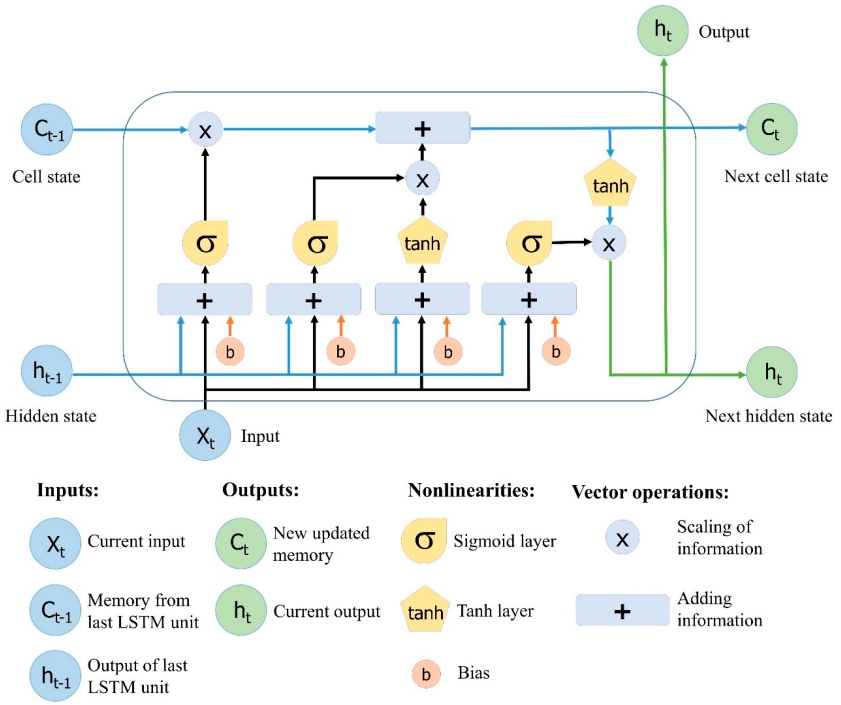


Figure: The long short-term memory (LSTM) architecture, Le et al. (2019).

Objective

- Building a global **hydrological model** using the **hybrid approach**
- **Proof-of-concept**, hybrid models of environmental systems not yet tested
- Retrieve **interpretable, data-driven estimates of the water cycle** states and fluxes

Data

Spatial

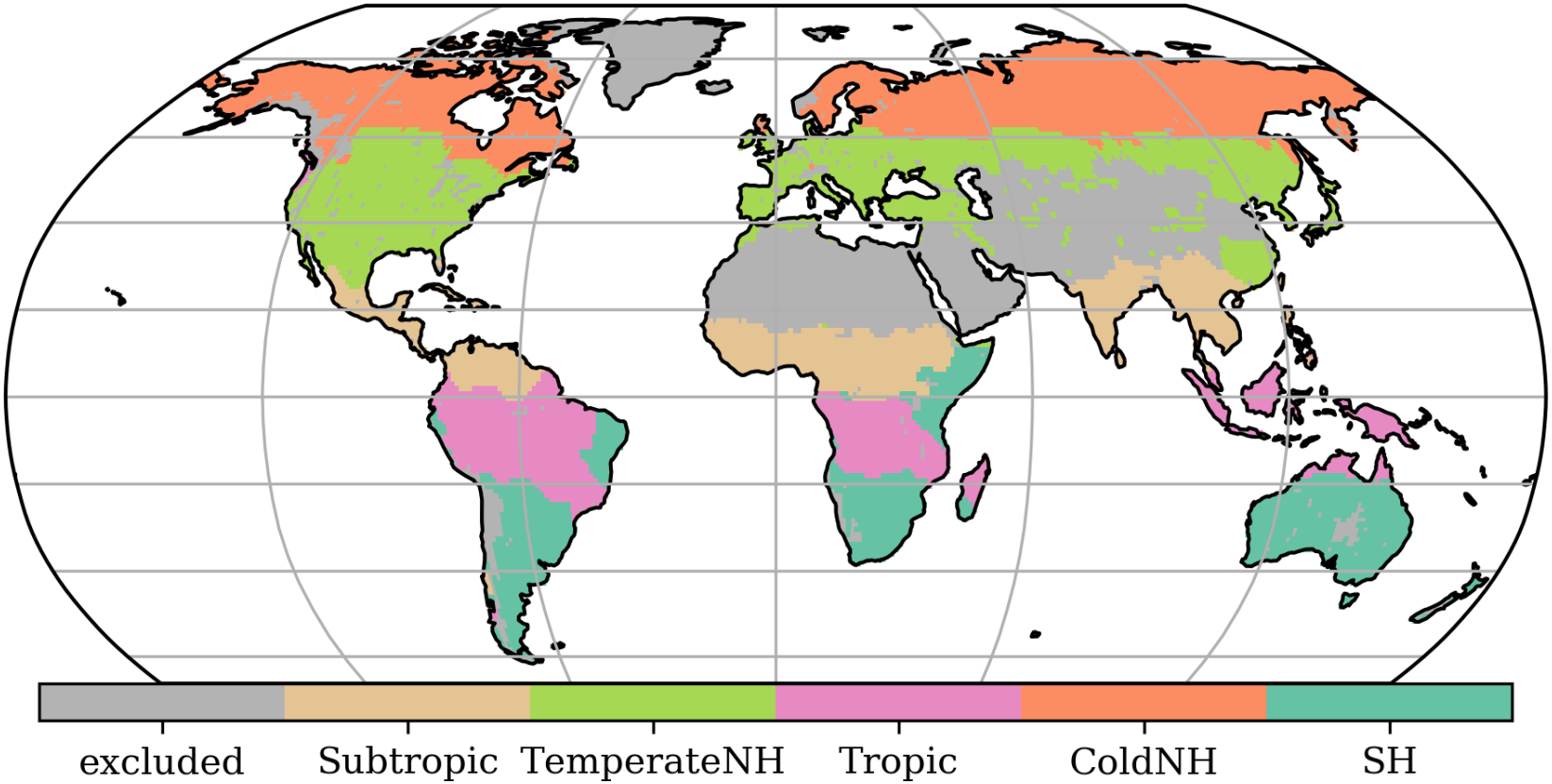


Figure: Spatial domain: global, 1.0°. De bioclimatic regions are used in the model evaluation.

Temporal

- Model runs on daily resolution, some target variables are on monthly resolution
- Training: 2002-01 to 2008-12
- Validation & test: 2009-01 to 2014-1

Datasets

Variable	Short	Dataset	Temporal	Spatial
Temporal features				
Precipitation	-	GPCP	Daily	1.0°
Net radiation	-	CERES	Daily	1.0°
Air temperature	-	CRUNCEP	Daily	1.0°
Static features				
Soil properties	-	Soilgrids	-	$\frac{1}{30}$°
Land cover fractions	-	Globland30	-	$\frac{1}{30}$°
Digital elevation model	-	GTOPO	-	$\frac{1}{30}$°
Wetlands	-	Tootchi	-	$\frac{1}{30}$°
Targets				
Total water storage	TWS	GRACE mascons	∼Monthly	1.0°
Evapotranspiration	ET	Fluxcom	Monthly	1.0°
Runoff	Q	GRUN	Monthly	1.0°
Snow water equivalent	SWE	Globsnow	Daily	1.0°

Table: Feature and target datasets.

Methods

The hybrid hydrological model

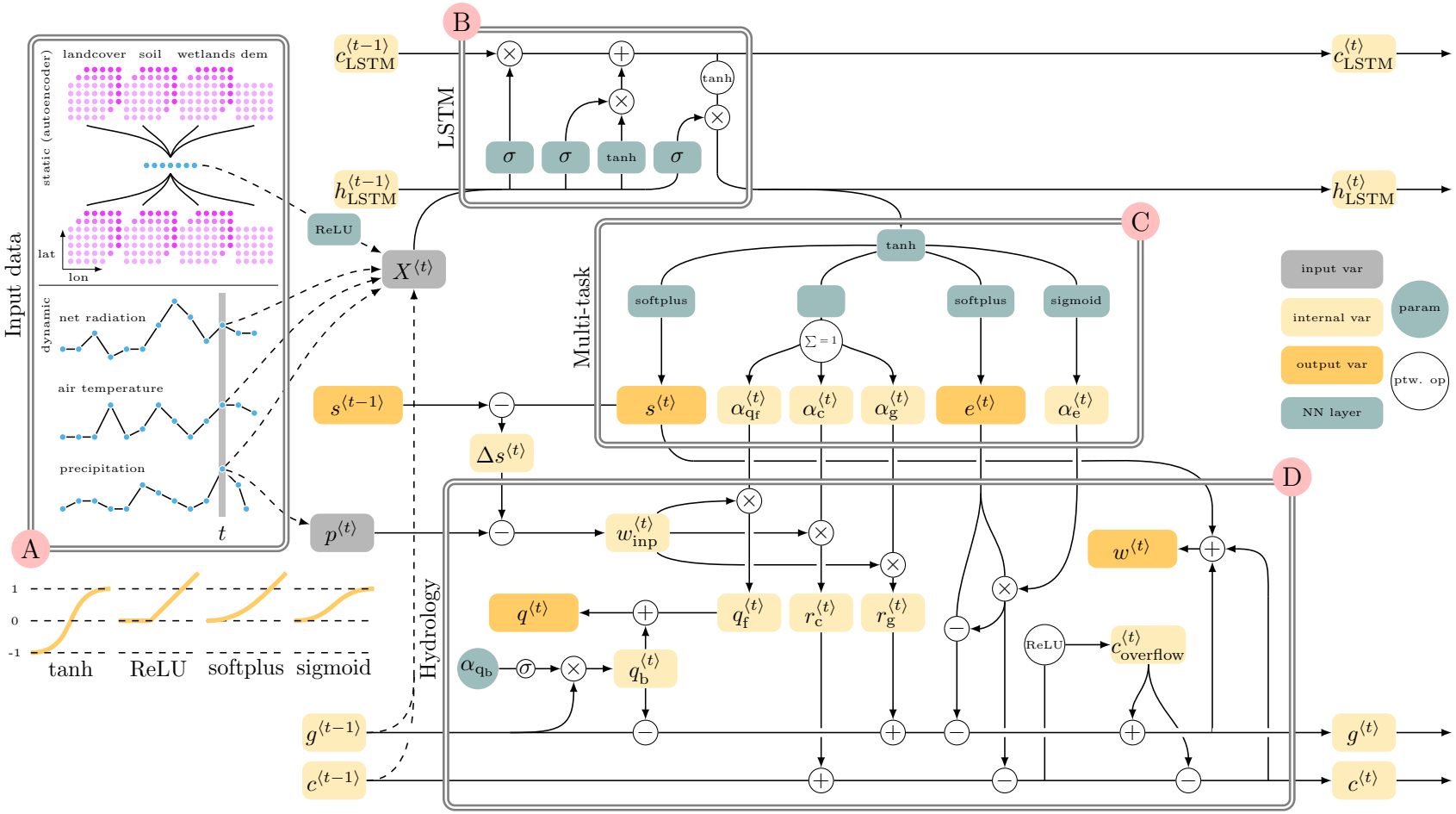


Figure: The proposed hybrid model, Kraft et al. (in preparation).

The hybrid hydrological model

Ⓐ **Input data:** the meteorological time series, encoded static variables and physically interpretable states groundwater (g) and cumulative water deficit (c) are fed into the LSTM. The dimensionality of the static variables is reduced in a preprocessing step using a convolutional autoencoder. The dimensionality is then further reduced using a feed-forward neural network with a bottleneck layer of size 12.

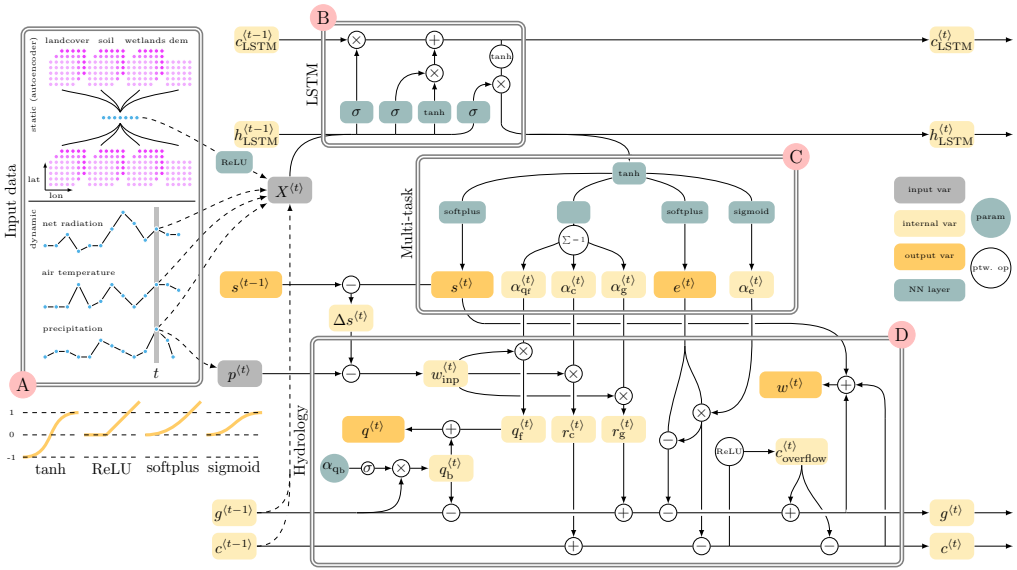


Figure: The proposed hybrid model, Kraft et al. (in preparation).

The hybrid hydrological model

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

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

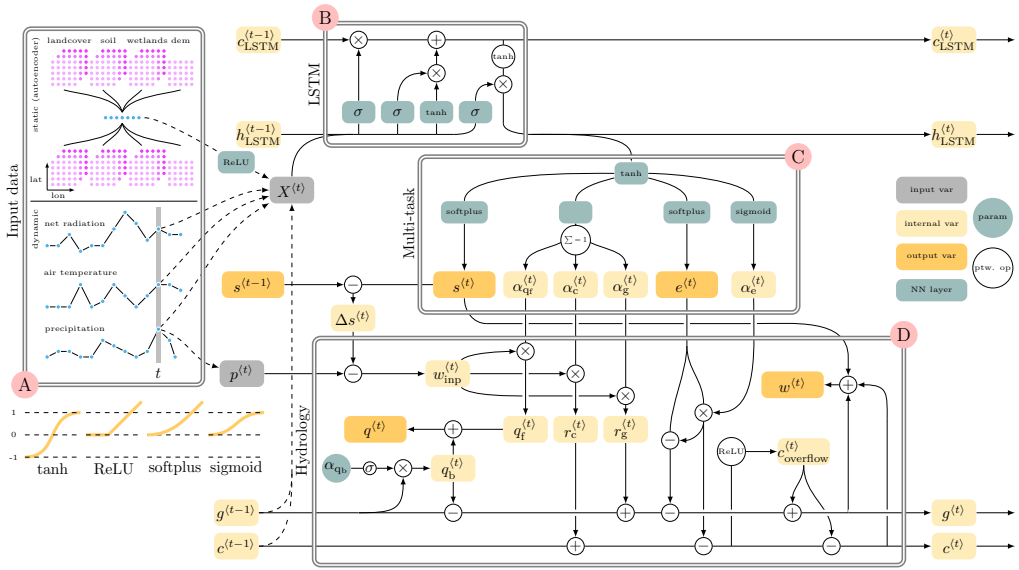


Figure: The proposed hybrid model, Kraft et al. (in preparation).

The hybrid hydrological model

C **The multi-task layer**, comprising of independent feed-forward layers (NN) yields interpretable variables: evapotranspiration (e), snow water equivalent (s), and fractions (α) defining how the liquid water input (w_{inp}) is partitioned into the fluxes of fast runoff (q_{f}), soil recharge (r_{c}) and groundwater recharge (r_{g}), as well as a fraction (α_{e}), that determines the source pool from which e is taken from. This mechanism compensates missing soil recharge from groundwater, which has been neglected to avoid equifinalities. The current w_{inp} is the precipitation (p) minus snow accumulation or plus snow melt (Δs).

$$e^{\langle t \rangle} = \text{softplus}(\text{NN}(h_{\text{LSTM}}^{\langle t \rangle}))$$

$$s^{\langle t \rangle} = \text{softplus}(\text{NN}(h_{\text{LSTM}}^{\langle t \rangle}))$$

$$\alpha_{\text{q}_\text{f}}^{\langle t \rangle}, \alpha_{\text{r}_\text{c}}^{\langle t \rangle}, \alpha_{\text{r}_\text{g}}^{\langle t \rangle} \overset{\sum=1}{=} \text{softplus}(\text{NN}(h_{\text{LSTM}}^{\langle t \rangle}))$$

$$\alpha_{\text{et}}^{\langle t \rangle} = \text{sigmoid}(\text{NN}(h_{\text{LSTM}}^{\langle t \rangle}))$$

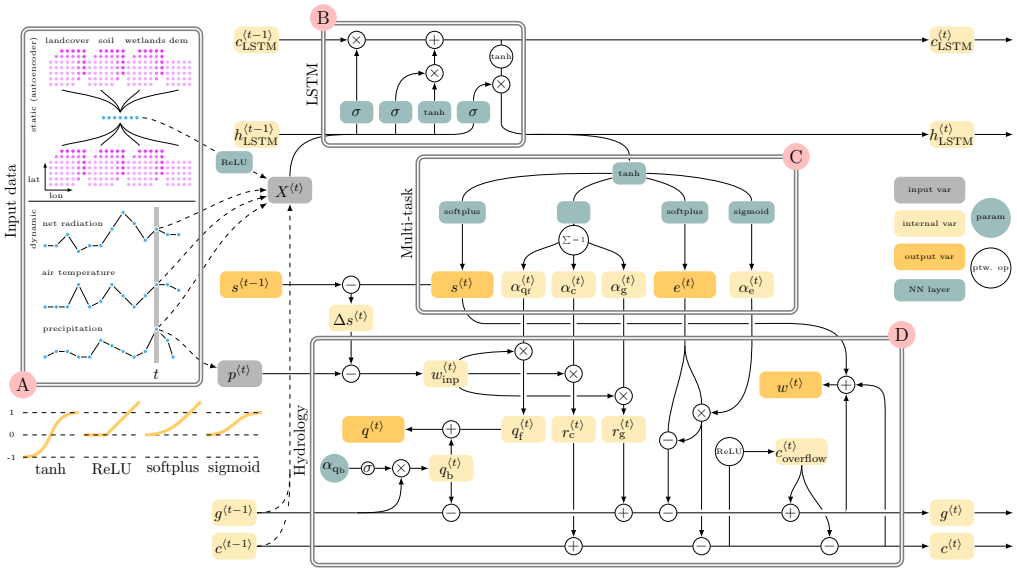


Figure: The proposed hybrid model, Kraft et al. (in preparation).

The hybrid hydrological model

$$c^{\langle t \rangle} = c^{\langle t-1 \rangle} + \overbrace{\alpha_{r_c}^{\langle t \rangle}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{r_c^{\langle t \rangle}} - e^{\langle t \rangle}\alpha_{e_c}^{\langle t \rangle}$$

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

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

$$g^{\langle t \rangle} = g^{\langle t-1 \rangle} - q_b^{\langle t \rangle} + \overbrace{\alpha_{r_g}^{\langle t \rangle}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{r_g^{\langle t \rangle}} + c_{\text{overflow}}^{\langle t \rangle} - e^{\langle t \rangle}(1 - \alpha_e^{\langle t \rangle})$$

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

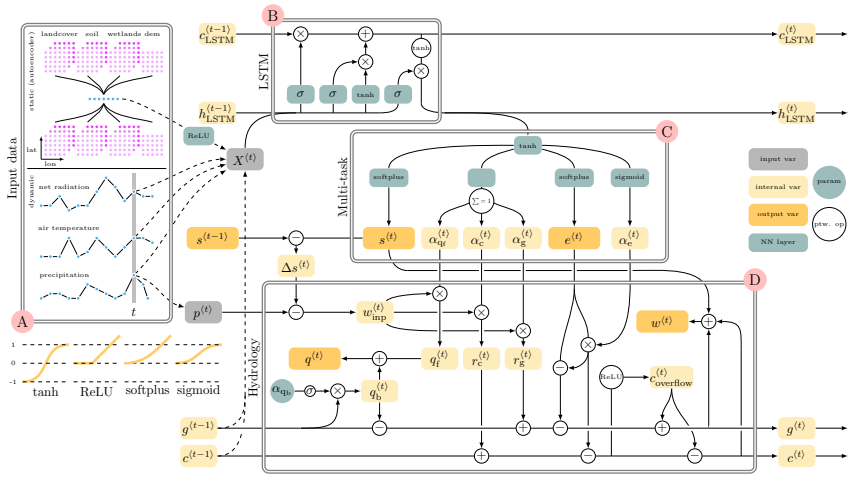


Figure: The proposed hybrid model, Kraft et al. (in preparation).

D **The water balance model** block implements hydrological balance equations. The physically interpretable state variables, g and c are updated at each time-step using a combination of the above latent variables and variables derived here. A learned global constant α_{q_b} is the fraction of g that is discharged as base runoff q_b. The total runoff q is the sum of q_b and q_f. The total water storage (w) anomalies are calculated as the sum of s, g, and c, minus the mean of w to get the variation around 0.

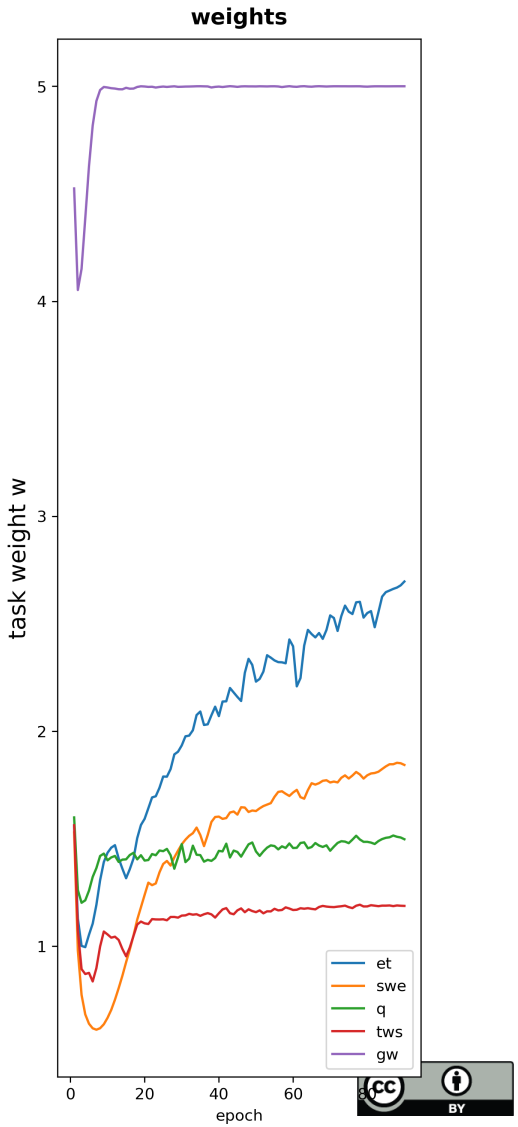
Self paced multitask learning

- Multiple target variables: TWS, SWE, ET and Q
- Instead of summing the loss terms, we use a weighted sum as proposed by Kendall et al. (2018).
- The task weights (σ) are parameters of the models that are updated dynamically.
- Further constraint added to penalize negative values for groundwater (GW).

Self paced multitask learning

$$loss = \sum_i^n \frac{1}{2*\sigma_i^2} loss_i + log(\sigma_i) = \sum_i^n w_i \, loss_i + r_i$$

where w_i is a weight for the task i reciprocal to the task uncertainty σ_i and r_i is a regularization term to prevent the uncertainty from converging to infinity.



Cross-validation

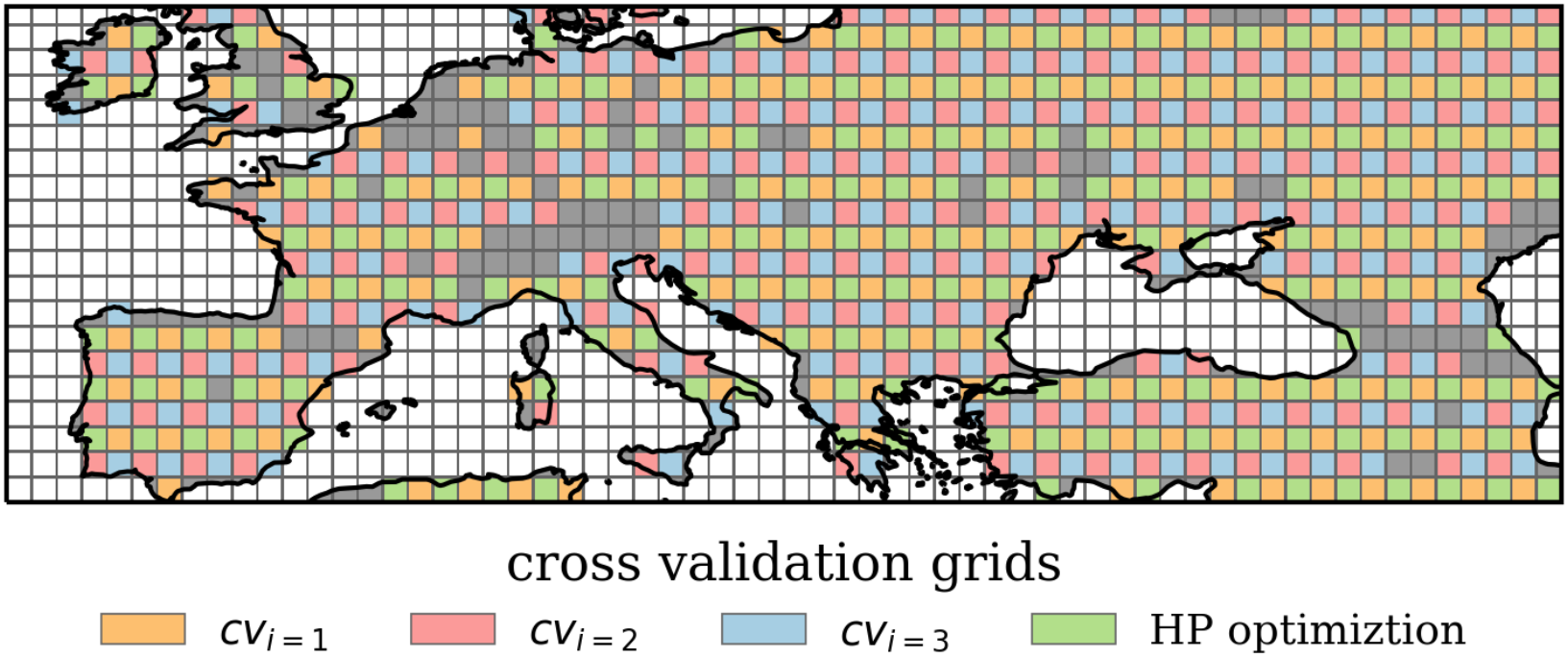


Figure: Spatial cross-validation scheme (subset shown): One grid is used for the hyper-parameter optimization (Bayesian optimization hyper-band (BOHB) algorithm Falkner et al. 2018), the other three grids for independent cross-validation runs. Each grid is split into five folds, one withheld for testing, the others are iterated such that each is used for validation once, the others for training.

Results

Model fit

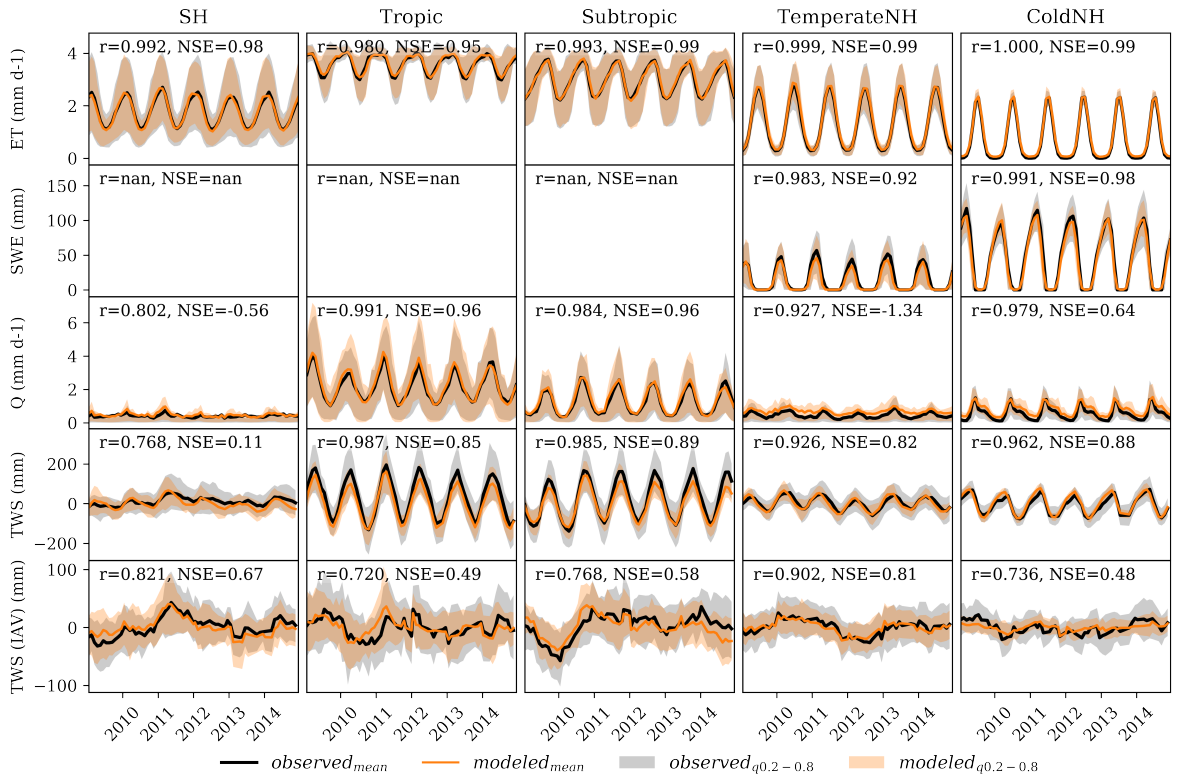


Figure: Test set model performance by bioclimatic regions. The target varialbes evapotranspiration (ET), snow water equivalent (SWE), runoff (Q), total water storage (TWS), and TWS interannual variability (IAV) are shown. The TWS IAV is the deviation from the mean seasonal cycle. Shaded areas are the $0.2 - 0.8$ quantiles of the regional variablity. Pearson correlation coefficient (r) and the Nash–Sutcliffe model efficiency coefficient (NSE) of the regional mean are shown.

Model fit

- Model fits target variables well (mostly NSE $>$0.8)
- Good fit (amplitude and timing) of TWS IAV (mostly NSE $>$0.5)
- ET & Q are upscaled products (machine learning based), thus probably easy to learn

Model robustness

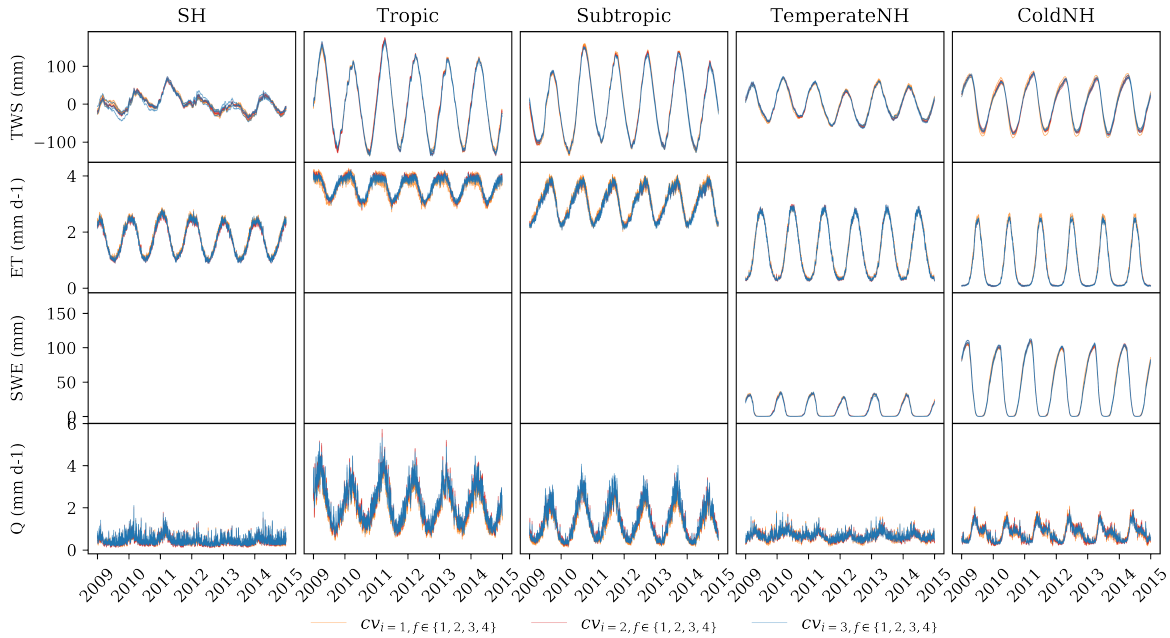


Figure: Model robustness w.r.t. target variables: regional mean of repeated model simulations of total water storage (TWS), evapotranspiration (ET), snow water equivalent (SWE), and runoff (Q). The lines represent the mean value of a single cross-validation. The lines are colored by cross-validation index *i*, i.e. lines with the same color come from one cross-validation run and represent the same grid-cells.

Model robustness

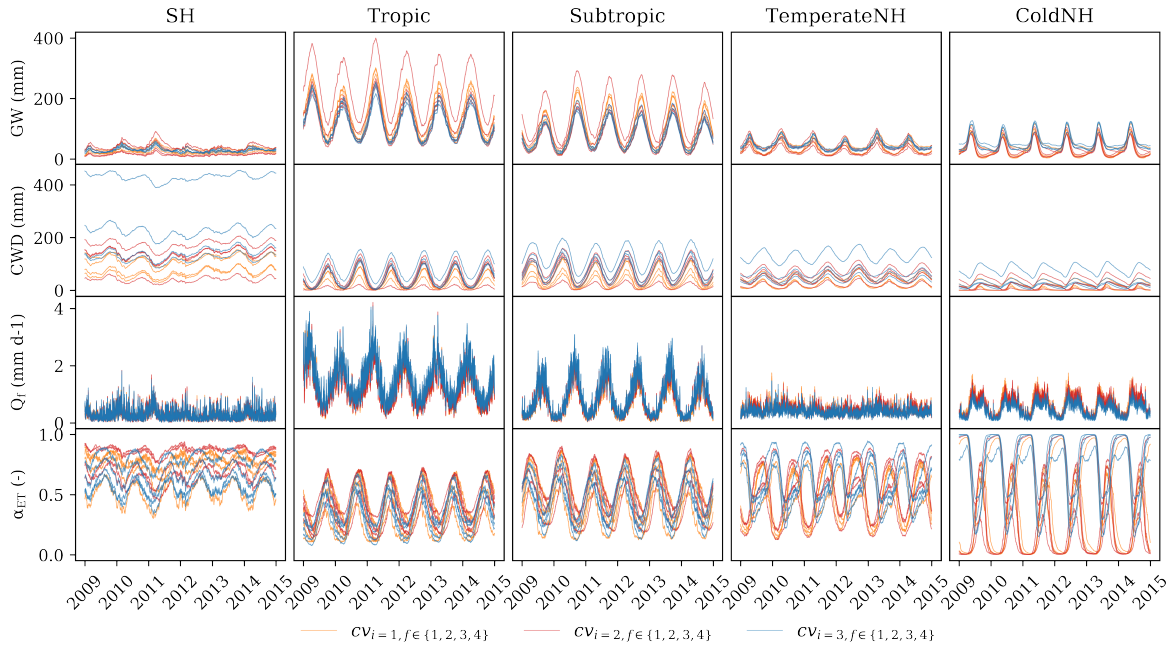


Figure: Model robustness w.r.t. latent variables: Regional mean of repeated model simulations of groundwater (GW), soil cumulative water deficit (CWD), fast runoff (Q_f), and frations (α_{et}), defining to what share evapotranspiration is extract form the soil versus groundwater. The lines represent the mean value of a single cross-validation. The lines are colored by cross-validation index i, i.e. lines with the same color come from one cross-validation run and represent the same grid-cells.

Model robustness

- Target variables: very robust
- Latent variables: varying robustness
- Under some conditions: underconstrained optimization problem

Complementary approaches

- In conceptual models: challenge to build a model that predicts target variables well
- Here: challenge to constrain model

Conclusion

Conclusion

- New data-driven approach to environmental modeling
- Good model fit but latent variables estimates of varying robustness
- Further constraines (through model revision, further soft constraints or further data constraints) needed

Conclusion

- We can learn from this model, even if imperfect (like conceptional models are as well)
- More data-driven perspective, interesting synergies with conceptual approaches

Literature

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.

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

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.

Kraft, B., Jung, M., Körner, M., and Reichstein, M. (in preparation). "HYBRID MODELING: FUSION OF A DEEP LEARNING APPROACH AND A PHYSICS-BASED MODEL FOR GLOBAL HYDROLOGICAL MODELING". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Le, X.-H., Ho, H. V., Lee, G., and Jung, S. (2019). "Application of long short-term memory (LSTM) neural network for flood forecasting". In: *Water* 11.7, p. 1387. DOI: doi.org/10.3390/w11071387.

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.

Samaniego, L., Kumar, R., and Attinger, S. (2010). "Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale". In: *Water Resources Research* 46.5. DOI: 10.1029/2008WR007327.