Robustness of Process-Based versus Data-Driven Modeling in Changing Climatic Conditions

SUNGMIN O

Department for Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany

EMANUEL DUTRA

Instituto Dom Luiz, Faculty of Sciences, University of Lisbon, Lisbon, Portugal

RENE ORTH

Department for Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany

(Manuscript received 16 March 2020, in final form 1 July 2020)

ABSTRACT

Future climate projections require Earth system models to simulate conditions outside their calibration range. It is therefore crucial to understand the applicability of such models and their modules under transient conditions. This study assesses the robustness of different types of models in terms of rainfall–runoff modeling under changing conditions. In particular, two process-based models and one data-driven model are considered: 1) the physically based land surface model of the European Centre for Medium-Range Weather Forecasts, 2) the conceptual Simple Water Balance Model, and 3) the Long Short-Term Memory-Based Runoff model. Using streamflow data from 161 catchments across Europe, a differential split-sample test is performed, i.e., models are calibrated within a reference period (e.g., wet years) and then evaluated during a climatically contrasting period (e.g., drier years). Models show overall performance loss, which generally increases the more conditions deviate from the reference climate. Further analysis reveals that the models have difficulties in capturing temporal shifts in the hydroclimate of the catchments, e.g., between energy- and water-limited conditions. Overall, relatively high robustness is demonstrated by the physically based model. This suggests that improvements of physics-based parameterizations can be a promising avenue toward reliable climate change simulations. Further, our study illustrates that comparison across process-based and data-driven models is challenging due to their different nature. While we find rather low robustness of the data-driven model in our particular split-sample setup, this must not apply generally; by contrast, such model schemes have great potential as they can learn diverse conditions from observed spatial and temporal variability both at the same time to yield robust performance.

1. Introduction

Land surface–hydrology models (LSMs) have been evaluated against observations and intercompared with each other over the past few decades, often by joint efforts of broad international groups, in order to identify strengths and inadequacies of existing model schemes. For instance, the Project for Intercomparison of Land Surface Parameterization Schemes (PILPS), launched in 1993, has examined differences among the participating models in the formulation of individual processes (Henderson-Sellers et al. 1993, 1995). The community effort was later expanded to regional and global scales (e.g., Dirmeyer et al. 1999; Boone et al. 2009; Dirmeyer 2011). PILPS also facilitated the Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER), which aims to reveal the potential for model improvements by benchmarking LSMs against simple linear regressions

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-20-0072.s1.

Corresponding author: Sungmin O, sungmino@bgc-jena.mpg.de

DOI: 10.1175/JHM-D-20-0072.1

(Best et al. 2015). Further recent multimodel experiments have focused on the role of LSMs in coupled climate models (e.g., Koster et al. 2006; Wei and Dirmeyer 2010; Seneviratne et al. 2013), some investigated model performance for hydrologic simulations (e.g., Reed et al. 2004; Haddeland et al. 2011; Beck et al. 2017), and others evaluated simulated energy fluxes and balances (e.g., Abramowitz et al. 2008; Jiménez et al. 2011). These experiments have led to a better understanding of land surface processes and a correspondingly more accurate representation within the models.

Meanwhile, with a growing recognition of nonstationarity in climate and hydrology, the functioning of models outside their training climate has become a focus of interest. Ongoing climate change challenges the reliability of LSMs for future climate projections, even if they function well under current conditions (Xu et al. 2005; Milly et al. 2008; Merz et al. 2011). As a result, numerous research studies have been undertaken by a range of model communities (so far, rather sporadically) to assess the extrapolation capacity of models under nonstationary conditions. To this end, corresponding model experiments have followed a differential split-sample testing (DSST; Klemeš 1986; Refsgaard et al. 2014), i.e., models are calibrated within a reference period (e.g., wet and cold seasons) and then assessed during a period characterized by different climate conditions (e.g., dry and warm seasons). Examples of application of DSST can be found in Seibert (2003), Wilby (2005), Vaze et al. (2010), Merz et al. (2011), Coron et al. (2012), Li et al. (2012), Seiller et al. (2012), Brigode et al. (2013), Kling et al. (2015), Li et al. (2015), Seiller et al. (2015), Thirel et al. (2015a), Broderick et al. (2016), Fowler et al. (2016), and Vormoor et al. (2018). However, no common method for testing models, e.g., selection of evaluation criteria, definition of conditions, has yet been agreed upon. Therefore, only preliminary conclusions could be drawn regarding model deficiency under changing conditions.

This study builds upon the earlier DSST studies for assessment of model robustness under changing conditions. Our model experiment focuses on rainfall–runoff modeling using data from 161 catchments in Europe encompassing diverse hydroclimatic regimes. Contrasting periods (i.e., wet versus dry conditions) are defined based on mean precipitation, because precipitation is the most commonly used climatic indication in the aforementioned DSST studies. For the first time, we extend the scope of such model evaluation by considering a diverse set of state-of-the-art models. Three different models with widely varying complexities are employed, namely, physically based, conceptual, and empirical models: the Hydrology-Tiled European Centre for Medium Range Weather Forecasting (ECMWF) Scheme for Surface Exchanges over Land (HTESSEL; Balsamo et al. 2009), the Simple Water Balance Model (SWBM; Koster and Mahanama 2012; Orth and Seneviratne 2015), and the Long Short-Term Memory-Based Runoff model (LSTM-Runoff, built in this study using LSTM; Hochreiter and Schmidhuber 1997), respectively. We define "model complexity" in terms of a degree of explicit consideration of physical knowledge or theoretical principles that govern hydrologic and relevant processes, rather than by the number of parameters or conceptual approaches. With this definition, HTESSEL is the most complex model, followed by SWBM and LSTM-Runoff. The LSTM-based machine learning is a relatively new approach in rainfall–runoff modeling, with a recently increasing interest because of its "memory"; the ability to store information from previous inputs for many time steps during model training. Its comparable performance to conventional models has been demonstrated, for instance, by Hu et al. (2018), Kratzert et al. (2018), Zhang et al. (2018), and Kratzert et al. (2019), yet, to our knowledge, there has been no hydrology study with a consideration on model transferability between contrasting conditions.

When it comes to process-based models, complex models tend to utilize a broader set of input data than simpler models and they are able to represent more processes and variables. However, often, conceptual models outperform their physically based counterparts in specific settings, and for particular variables, despite using less input information and weaker physical constraints (e.g., Perrin et al. 2001; Materia et al. 2010; Orth et al. 2015; Tegegne et al. 2017). This can be explained, on the one hand, by higher calibration flexibility in simple models, bearing the risk of overfitting. On the other hand, complex models may suffer from an incomplete representation of land surface processes due to knowledge gaps (Beck et al. 2017) or from an incompatibility between relevant simulated processes (Koster and Milly 1997). Beven (1989) suggested the use of three to five parameters to reproduce the most dominant hydrologic processes while avoiding overparameterization. However, those studies did not explicitly consider the model skills in the context of changing conditions, but rather focused on the capacity of model to describe "current" processes. Some previous DSST studies have attempted to take into account model complexity (e.g., Vaze et al. 2010; Coron et al. 2012), but no significant difference between models was observed. It should be noted that in those DSST studies, the complexity was defined by the number of parameters or model structure among considered conceptual models.

Expanding upon known differences in the nature of conceptual and physically based models, we aim to test

TABLE 1. Models used in the study.

Model	Description	Type	Method	Complexity	Reference
HTESSEL	ECMWF land surface scheme	Physically based	Differential equations	High	Balsamo et al. (2009, 2015)
SWBM	Simple Water Balance Model	Conceptual	Simplified equations	Medium	Orth and Seneviratne (2015)
LSTM-Runoff	LSTM-based runoff model	Data-driven	Black-box concept	Low	Hochreiter and Schmidhuber (1997)

in this study to what extent model complexity influences the robustness of model performance under transient climate conditions. More complex models typically represent a wider range of relevant land surface processes. The physically based model HTESSEL also implements land energy processes, while the conceptual model SWBM describes only hydrologic processes. Further, the LSTM-Runoff is designed to merely produce runoff time series. All three models are calibrated or trained using streamflow observations only. We also seek to elucidate how different hydroclimatic conditions between contrasting periods can influence model performance and discuss opportunities for enhancement of model robustness under transient conditions. Additionally, we test our results with different, equifinal sets of well-performing parameters to analyze the impact of parameter selection on our robustness assessment. The following section 2 and section 3 describe models and methodology used in this study, respectively. The results are presented in section 4. Discussion and conclusions are summarized in section 5.

2. Models

In this study, HTESSEL, SWBM, and LSTM-Runoff represent physically based, conceptual, and empirical models, respectively. Note that both HTESSEL and SWBM include representations of hydrologic processes and are therefore referred to as ‘‘process-based models’’ hereafter. In contrast, LSTM-Runoff contains no explicit physical or conceptual representation of the hydrologic processes. The model can self-learn the relation between input and output, making it a ‘‘data-driven model.’’ The main characteristics of each model are summarized in Table 1.

a. Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land

HTESSEL is a land surface model used operationally in the Integrated Forecast System (IFS) of ECMWF for short-range forecasts to seasonal predictions, and reanalysis, to simulate surface water and energy fluxes and the evolution of soil and snow (Balsamo et al. 2009). A grid box is divided into up to six land tiles representing different subgrid surface types. In each grid box, two vegetation types (a high and a low vegetation) are represented. Vegetation growth and decay is seasonally variable, but does not respond to weather and climate anomalies. An interception layer accumulates precipitation until it is saturated, and excess precipitation is partitioned into surface runoff and infiltration. The subsurface fluxes are modeled in four layers and a single layer snowpack. More detailed information can be found in ECMWF (2016) and Balsamo et al. (2015).

b. Simple Water Balance Model

SWBM is a conceptual, lumped model originally proposed by Koster and Mahanama (2012) and has been modified and widely applied over European catchments by Orth and Seneviratne (2013, 2015). The model assumes simple dependencies of evapotranspiration (normalized by net radiation) and runoff (normalized by precipitation) on soil moisture. Runoff depends on precipitation and soil moisture only. The model accounts for snow with a degree-day method. In contrast to HTESSEL, subsurface flow is disregarded (bucket-type approach) and soil or vegetation information is not employed. More detailed information can be found in Orth and Seneviratne (2013).

c. Long Short-Term Memory-Based Runoff model

LSTM is a specific type of recurrent neural network that was designed to model sequences (e.g., time series) and their long-term dependencies (Hochreiter and Schmidhuber 1997). Recently, LSTM-based approaches are increasingly used in Earth system science including hydrologic modeling (e.g., Hu et al. 2018; Kratzert et al. 2018; Zhang et al. 2018; Kratzert et al. 2019; Sahoo et al. 2019). In this study, LSTM-Runoff is designed to predict runoff at a given time step using meteorological forcing over multiple prior time steps (look-back). Note that the model has no knowledge of land surface or hydrologic processes. LSTM-Runoff itself learns a relation between input and output series (meteorological conditions and runoff in our case) by repeated updating of trainable parameters (weights), followed by the computation of error signals for all parameters; i.e., feed-forward and back-propagation. On the other hand, users specify a model configuration through hyperparameters which determine model architectures and learning methods. For instance,

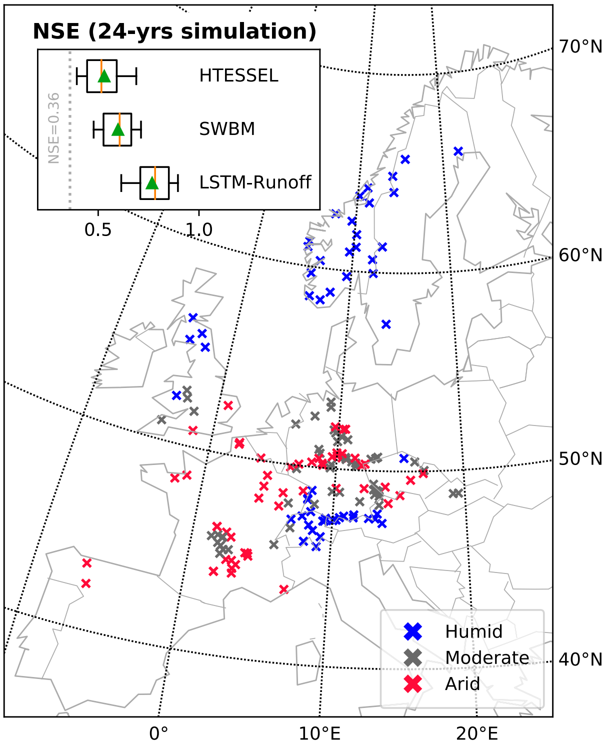

FIG. 1. Location of the 161 study catchments distributed across Europe. Inset shows the summary of NSE values for model simulations over the entire study period; 25th and 75th percentiles (boxes), median (lines within boxes), range from 10th to 90th percentiles (whiskers), and mean (triangles). The catchments are grouped into humid, moderate, and arid regions according to their aridity index. The map is created using the Matplotlib basemap toolkit (Hunter 2007).

the architecture-related hyperparameters include input dimension, the amount of neurons, and the number of hidden layers, while the learning-related hyperparameters include cost functions, iterations of forward and backward propagation for all training data (epochs), and dropping out units as a regularization technique (dropout).

3. Methodology

a. Study area

In an initial setup, 398 near-natural catchments over Europe are considered, where long-term streamflow records (mostly from 1984 to 2007) are available (Stahl et al. 2010). To ensure the general applicability of models, we select a subset of catchments in which all three models exhibit satisfactory performance; see section 3c for more details. Eventually, a total of 161 catchments in 11 countries is chosen (Fig. 1). The median basin size is 261 km^2, ranging from 4 to 3781 km^2. For further analysis, the chosen catchments are then grouped into three different hydroclimatic regimes according to their aridity index (long-term dryness; Budyko 1974): humid, moderate, and arid. Aridity index, the ratio of atmospheric water supply to demand, is defined as the ratio between mean net radiation and respective unit-scaled precipitation during the entire 24 years. The grouping indicates only relative dryness among the catchments because the same number of catchments are assigned to each group using the 33.3rd and 66.6th percentiles as thresholds.

b. Forcing data

The models are all driven in an uncoupled mode with daily forcing data at a 0.5° × 0.5° scale and applied in a lumped fashion. All models commonly employ precipitation, temperature, and radiation information, while HTESSEL and LSTM-Runoff further use additional forcing variables (Table 2). Forcing datasets are obtained from the WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset (Weedon et al. 2014), except for precipitation and temperature data which are obtained from the station-based E-OBS dataset (Cornes et al. 2018). Precipitation data are upscaled by 10% to account for undercatch biases (Hofstra et al. 2009), following Orth and Seneviratne (2015). Net radiation data for SWBM are obtained from ERA-Interim (Dee et al. 2011). For HTESSEL, precipitation data are preprocessed to be assigned as rainfall when 2-m temperature exceeds 0°C, and solid phase otherwise. In this study, LSTM-Runoff is trained using the same forcing data (features) as for HTESSEL to derive the main results. It is further trained with the same inputs as for SWBM in the context of the hyperparameter uncertainty analysis in section 4c. In the case of LSTM-Runoff, precipitation is log-transformed. In addition, all input data are normalized using their mean and standard deviation for training efficiency (LeCun et al. 2012). For HTESSEL, static information describing land cover (vegetation cover and types), soil textures and mean climatologies of leaf area index and surface albedo are also required. No climatic or static data are used in SWBM or LSTM-Runoff.

c. Simulation setup

This section outlines the simulations setup implemented in this study (see also Table 3). First, we carry out continuous simulations over the entire time period (i.e., 24 years) to test whether models adequately reproduce runoff in each catchment. Second, models are calibrated for each catchment during reference periods, except for LSTM-Runoff which is trained using data from the reference periods of all catchments at once, as further explained in section 3c(2). This is done for the

TABLE 2. Daily atmospheric forcing data (1984–2007) at 0.5° spatial resolution. An x indicates that the data are used for models.

Variable	Data source	HTESSEL	SWBM	LSTM-Runoff
Surface incident longwave radiation	WFDEI	x		x
Surface incident shortwave radiation	WFDEI	x		x
Net radiation	ERA-Interim		x	
Precipitation	E-OBS	x	x	x
Temperature	E-OBS	x	x	x
Near-surface specific humidity	WFDEI	x		x
Surface pressure	WFDEI	x		x
Wind speed	WFDEI	x		x

wettest and driest years, respectively. Finally, models are run over all remaining years so that model results can be examined along the changing conditions, i.e., across increasingly drier years (wet2dry) and increasingly wetter years (dry2wet), respectively.

1) MODEL SIMULATIONS FOR CATCHMENT SELECTION

We run the models over the entire time period for each catchment to verify the general suitability of the models and the quality of input–output data. In the case of the process-based models, a total of 500 simulations are carried out for every catchment, each with a different, randomly generated parameter set. Information about model calibration and parameter sampling can be found in appendix. When any of the 500 simulations yields a Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970) that exceeds or equals to 0.36, for both HTESSEL and SWBM, the catchment is selected for further study. The NSE criterion is adopted from Motovilov et al. (1999) and Moriasi et al. (2007). As a result, 161 suitable catchments are selected. Model efficiency for the selected catchments is shown in the inset of Fig. 1. The remaining catchments, where the NSE $<$ 0.36 for all 500 simulations by either HTESSEL or SWBM, are discarded. The poor performance could be due to contaminated observations, e.g., by human interference, spatial mismatch of observations and model outputs, or shortcomings of the models themselves. Note that the performance of LSTM-Runoff is not considered here to select the study catchments. This is because the model could yield reasonably high performance, by an undesired overfitting, with all catchment observations given its large number of parameters and flexibility which is not limited by physical constraints. For verification purposes, the performance of LSTM-Runoff in the above-selected 161 catchments is determined, showing consistently NSE $\geq$ 0.36 with epochs = 50, as expected.

2) MODEL CALIBRATION TO REFERENCE CONDITIONS

For each catchment, we use the wettest calendar year for model calibration and retain the remaining 23 years as an independent evaluation period. The same is carried out vice versa with the driest calendar year. The wet/dry conditions are defined according to annual mean precipitation. Note that even during the driest year few runoff peaks may occur which then inform model calibration. Precipitation is the most commonly used climatic characteristic in DSST as a primary driver of the natural hydrologic cycle, and consequently runoff. Moreover, Coron et al. (2012) showed that precipitation is more influential to the dependency of model performance (runoff simulation) on considered conditions than other climatic variables like temperature or potential evapotranspiration. For the process-based models, the 500 parameter sets sampled in the previous section are used

TABLE 3. Model simulations.

Simulation	Time period	No. of catchments	Purpose	Used (hyper)parameters
Catchment selection	Complete period (24 years)	398	Check general model suitability and data quality in tested catchments	500 parameter sets for HTESSEL and SWBM, while hyperparameters of interest for LSTM-Runoff
Calibration	The wettest/driest year, while randomly chosen year for LSTM-Runoff*	161	Train models under hypothetical reference conditions	500 parameter sets for HTESSEL and SWBM, while hyperparameters of interest for LSTM-Runoff
Evaluation	Remaining drier/wetter years (mostly 23 years)	161	Evaluate model performance under contrasted conditions	Best-performing (hyper)parameter set during calibration

here again to calibrate the models for each of the 161 catchments. Given our calibration with one year of forcing data, the model simulations are iterated repeatedly over the reference year until the model state reaches an equilibrium, i.e., model output values remain unchanged.

Even when calibrated with intentionally constrained training data, the process-based models can benefit from their physics foundations given that these foundations are based on knowledge from diverse climate conditions. The data-driven model does not rely on such foundations but learns the input–output relationship exclusively from the training data. Consequently, setting up the split-sample experiment, in a way that is comparable across all the different model types, is not straightforward. In any case, a fair setup of the data-driven model requires more (diverse) training data than for the other two models to compensate for the missing physics foundations, which represent knowledge from diverse climate conditions. Therefore, we proceed with two versions of the data-driven model, also to illustrate the impact of the training data and strategy. First, LSTM-Runoff is trained with the reference years from all 161 catchments (=365 days × 161 days $\simeq$ 60 000 days). In this way, LSTM-Runoff is trained with more (hydroclimatically diverse) data than the process-based models (see Fig. S1 in the online supplemental material). Second, LSTM-Runoff* is trained with the same amount of data (~60 000 days) as LSTM-Runoff, but with one randomly selected year from each catchment rather than the respective extreme reference year. Therefore, LSTM-Runoff* is allowed to experience an even wider range of hydroclimatic conditions. Consequently, in this version the data-driven model can build its own knowledge from *all* observed conditions. The hyperparameters of LSTM-Runoff and LSTM-Runoff* are selected through a grid search with tenfold cross validation (see also appendix).

3) Model evaluation over years under changing conditions

With optimized parameters obtained from calibration over the reference periods, model simulations are carried out over the remaining years. The evaluation period is assumed to be representative of transient climatic conditions. This permits assessment of model performance along gradually changing conditions and thus inferring the respective behavior of the models. The process-based models are run repeatedly over each year until the models approach their equilibrium, just as conducted for calibration. For LSTM-Runoff, 10 runs for the selected network configuration are performed and final runoff is computed as an average of the 10 runs given the random initialization of the model; the model weights are initialized with the Glorot/Xavier initialization (Glorot and Bengio 2010). For Fig. 5 and the supplemental material, we perform five runs; the number of runs only marginally affects our results as we examine the average behavior of the model across catchments.

4. Results

In this section, we first assess the robustness of runoff simulations under changing conditions (i.e., wet2dry and dry2wet) and compare results between models and between aridity regimes. Second, model performance is further investigated along temporal shifts in aridity, in addition to changes in precipitation; the yearly aridity is considered as a potential factor affecting model performance in transient conditions. Finally, parameter uncertainty is assessed to test the role of the chosen (hyper)parameter sets for the conclusions of our study.

a. *Model robustness with respect to transient conditions*

We compare the three models, examining their performance along changing annual precipitation totals as a function of 1) NSE of daily runoff and 2) percentage bias of annual runoff Q_{diff}. With the optimal value of 0, positive values of Q_{diff} indicate an overestimation bias in simulated runoff, and negative values indicate underestimation. Figure 2 portrays the changing performance of the models in response to changing precipitation averaged across catchments (median). Given that the wettest/driest years are defined for each catchment, the number of performance values (NSE or Q_{diff}) at the reference period is the total number of catchments, i.e., 161 values (top panel of Fig. 2). Most, but not necessarily all, catchments contribute to all the following precipitation P bins over the validation period. If a catchment contributes more than one year for the same P bin, catchment-averaged performance is calculated first.

For wet2dry (Fig. 2, left), LSTM-Runoff and SWBM show comparable performance overall in terms of NSE (0.62 and 0.64, respectively) during the wettest years, while HTESSEL shows comparatively weaker performance (0.47). Similar results are noted for the driest years (Fig. 2, right); NSE values of 0.66, 0.61, and 0.42 for LSTM-Runoff, SWBM, and HTESSEL, respectively. Despite its higher complexity, the physically based model does not necessarily outperform the simpler model types, e.g., due to a lack of sufficient data to adequately characterize the model parameters. In turn, the physical constraints are weaker in LSTM-Runoff and SWBM such that these models can be forced to describe runoff hydrographs more accurately during the calibration period, which at the same time increases the potential risk of overfitting. For performance evaluation,

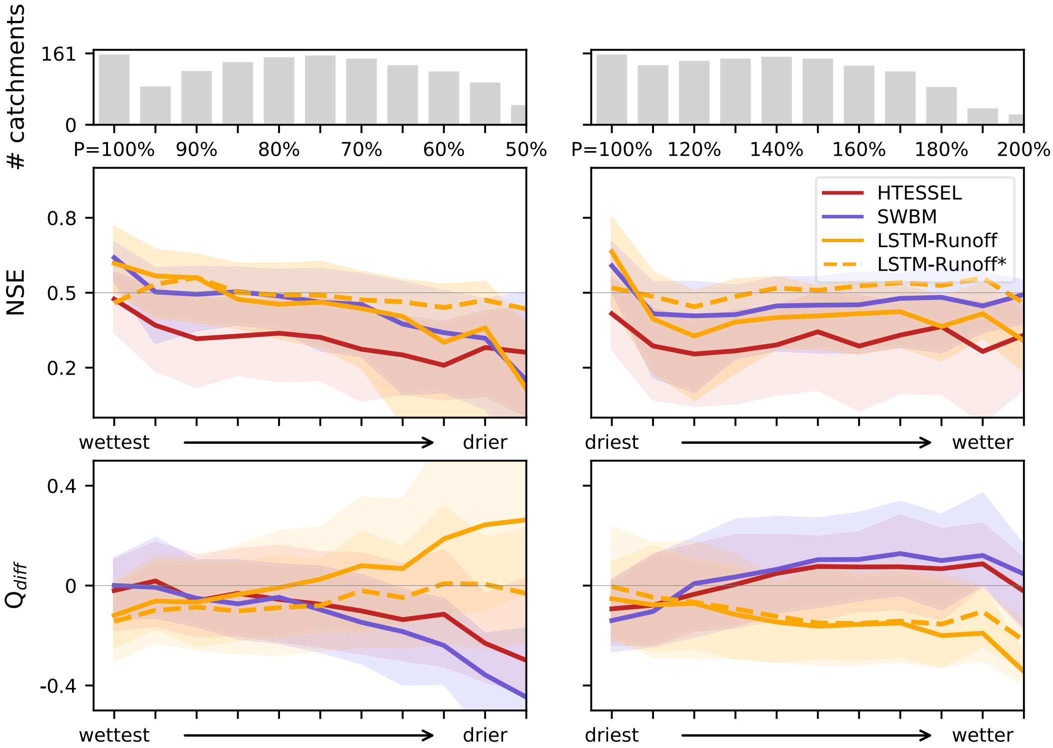

FIG. 2. Model performance under changing annual precipitation totals, (left) from wettest to drier and (right) from driest to wetter years. (top) The number of catchments that have at least 1-yr time series for each relative precipitation bin; total precipitation of the wettest or driest years is assumed to be 100%. (middle) NSE and (bottom) Q_{diff} (relative bias in annual runoff) are used to compare the performance between the models. Catchment-averages are calculated first if a catchment has more than 1 year for the same precipitation bin, before averaging across all catchments contributing data to a particular bin. Lines show median and the shading denotes the interquartile range.

runoff simulations from March to December are considered for all models in the figure as LSTM-Runoff simulations can only start in March due to the applied 60 days look-back (see section 2c and appendix). The performance of process-based models has been recalculated with outputs of the full 1-yr period and results remain very similar (not shown).

In the wet2dry simulation, all three models yield robust performance under conditions similar to the reference periods, i.e., $P > 70\%$. However, as the conditions change, an overall decrease in general performance, with respect to both NSE and Q_{diff}, is observed by all three models. The decline of NSE is faster for SWBM and LSTM-Runoff such that their performance becomes poorer than that of the physically based model as precipitation is reduced. Interestingly, model performance does not decrease at a constant rate, but there is an apparent acceleration at around $P = 70\%$. It is more clearly detected in the case of SWBM and LSTM-Runoff. A similar pattern is found for runoff biases. For dry2wet, we observe the opposite behavior with a rapid performance drop until approximately $P = 120\%$ and a levelling off afterward. These patterns of model performance deterioration are further analyzed in section 4b.

In the case of the Q_{diff} evolution across changing precipitation conditions, the process-based models and the data-driven model reflect opposite behaviors; increasing underestimation bias in HTESSEL and SWBM versus growing overestimation in LSTM-Runoff for wet2dry, and vice versa in the case of dry2wet. LSTM-Runoff probably adapts to the range of runoff during the training years. For instance, with a runoff ranging between 10 and 20 mm for the wettest year, LSTM-Runoff tends to predict runoff within this range even though the actual runoff range decreased (e.g., 5–10 mm) during the drier years, which leads to the overestimation bias. This explanation applies correspondingly for dry2wet. In the case of the process-based models, we hypothesize that the partitioning of precipitation into runoff and evapotranspiration (ET) is not flexible enough but rather fixed during calibration and wrongly maintained under changed conditions, leading to the observed runoff biases.

We also repeat the entire analysis by calibrating and evaluating model simulations with 2-yr periods instead

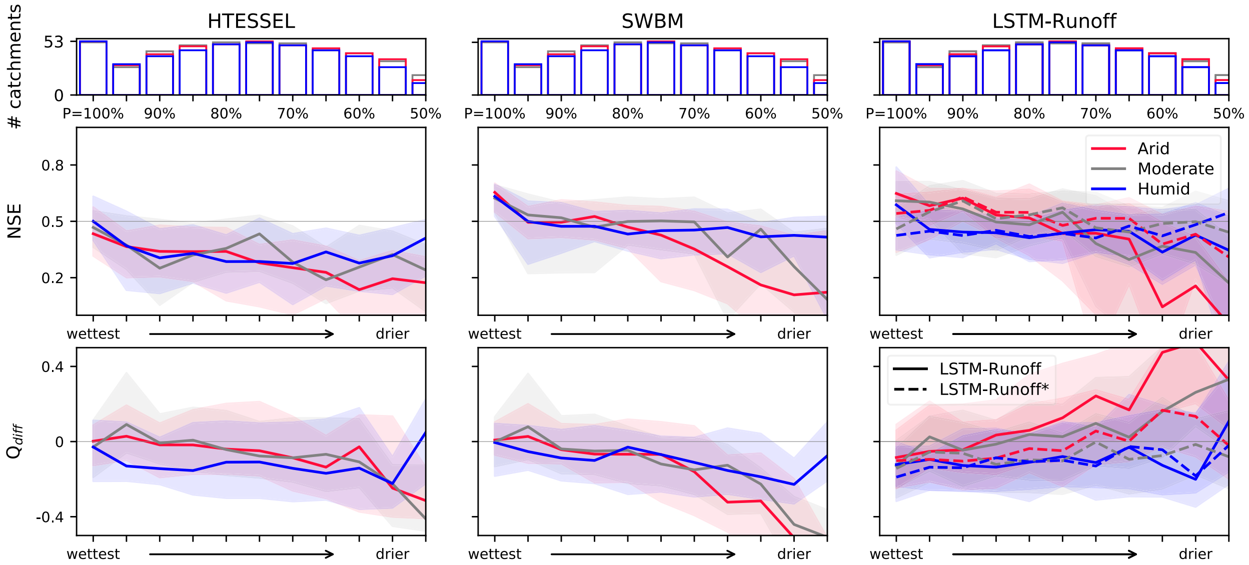

FIG. 3. As in Fig. 2, but for the catchments of humid, moderate, and arid groups (blue, gray, and red, respectively). Note that for simplicity only wet2dry results are shown. Results from dry2wet can be found in the supplemental material (Fig. S3).

of single years. Overall, we find similar results in terms of the difference between the changes in model performance, as displayed in Fig. S2. This illustrates that the use of 1-yr periods is not detrimental for model calibration.

When it comes to LSTM-Runoff*, better performance at the wettest or driest years is not clearly captured because the newly selected reference years for training the model are randomly distributed over the P bins. Both LSTM-Runoff perform almost equally well under conditions closer to the wettest/driest years. However, LSTM-Runoff* shows a stable performance along the changing conditions. It eventually outperforms LSTM-Runoff after the $P = 70\%$ in the case of wet2dry, despite the same amount of training data. It demonstrates that LSTM-Runoff* learns possible changes to a catchment from others' current conditions. The results stay the same when LSTM-Runoff* is trained with different combinations of randomly selected years (not shown).

The interquartile error range increases the more conditions deviate from the reference conditions, implying varying patterns of deterioration of model performance among the catchments. We evaluate model performance for each catchment group, defined in section 3a: humid, moderate, and arid regions. For brevity, only wet2dry results are discussed here. Model performance under the reference conditions is comparable among the region groups for the same models (Fig. 3). However, it is apparent that all three models show a faster decline in performance along the changing conditions in the arid group of catchments. On the contrary, the model performance remains relatively stable during the evaluation period in the more humid catchments. As expected, LSTM-Runoff* shows a highly robust performance for all catchment groups. Similar results are obtained for dry2wet and displayed in Fig. S3. The difference in the model robustness between the regions is further discussed in the following section.

b. *Why and when does model performance deteriorate?*

In this section, we introduce yearly aridity to understand observed changes in model performance in more details. The definition of yearly aridity is similar to the aridity index used for the catchment grouping; i.e., the ratio of mean net radiation to mean unit-scaled precipitation, but for each year. Obviously, catchment-averaged aridity is increasing for wet2dry due to both decreases in precipitation and increases in net radiation under drier conditions (Fig. 4, left). Aridity is an indicator for distinguishing water-limited versus energy-limited environments (Denissen et al. 2020). Therefore, a shift of yearly aridity across 1 implies profound changes in the hydroclimatic conditions experienced in the catchments.

The relative precipitation at which yearly aridity crosses 1 is found at $P \simeq 70\%$ for wet2dry and $P \simeq 120\%$ in the case of dry2wet. Interestingly, therefore the switch between water- and energy-limited conditions is found just where the models exhibit the aforementioned accelerated performance deterioration (Fig. 2). This is an important finding, as it presents a mechanistic explanation of model performance loss in transient climate conditions. Models apparently have difficulties in representing hydroclimatic regime shifts; in energy-limited

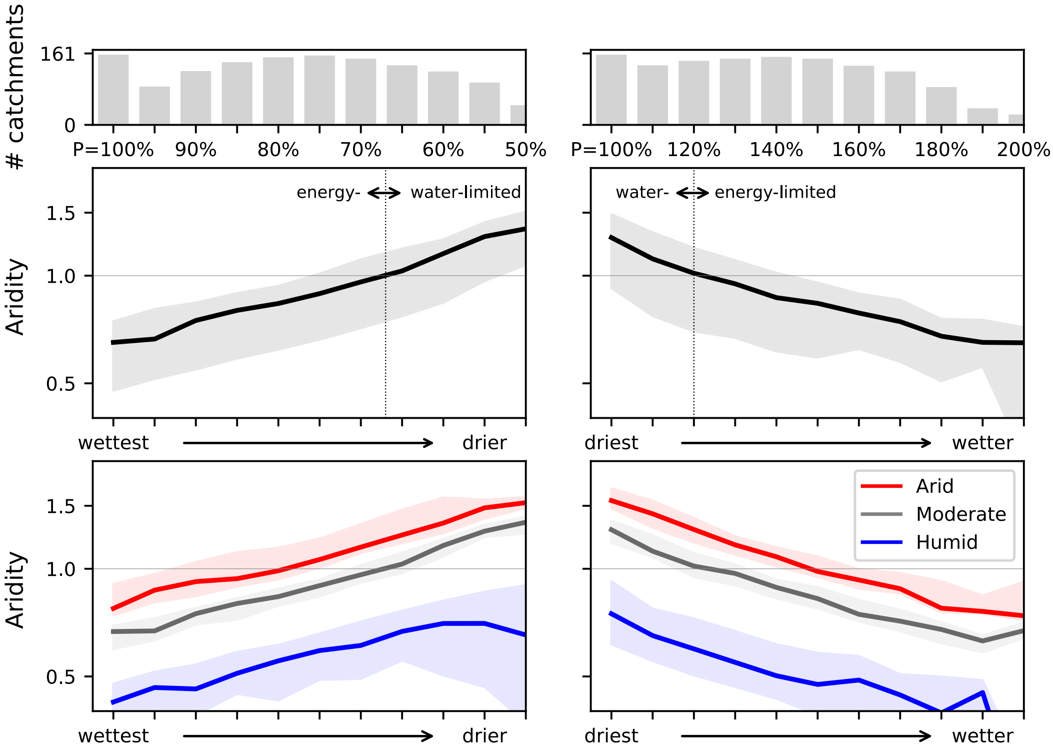

FIG. 4. Annual aridity along the changing annual precipitation totals, (left) from wettest to drier and (right) from driest to wetter years; averaged (middle) for all catchments and (bottom) for each catchment group. (top) The number of catchments for each relative precipitation bin. Lines show the median, and the shading denotes the interquartile range. Note the log scale of y axis.

environments, ET and hence surface temperature and moisture supply into the atmosphere, are controlled by atmospheric energy supply, while under water-limited environments this is mostly governed by soil moisture availability (Troch et al. 2009; Orth and Destouni 2018).

The temporal shifts in aridity also explain why we observe differences in model robustness between the groups of humid, moderate, and arid. Only aridity lines of the arid and moderate groups are crossing aridity = 1, meaning that the catchments of those groups are experiencing the regime shift for both wet2dry and dry2wet (Fig. 4, bottom). To the contrary, for the humid group the yearly aridity values are always found to be <1, allowing the models to deliver more reliable simulations even under different precipitation conditions. Also note that runoff modeling in water-limited environments (arid and semiarid) could be inherently more challenging owing to their distinctive hydrologic processes, e.g., absence of baseflow and bigger role of vegetation in hydrology (Pilgrim et al. 1988). Further, the yearly aridity of the reference years is closer to 1 in dry2wet than in wet2dry. This means that in the dry2wet case the models benefit from training data which more abundantly covers both water- and energy-limited environments (see also Fig. S1). This explains why the changes in model performance outside training conditions are less evident for dry2wet (Fig. 2, right). This is particularly the case for LSTM-Runoff and LSTM-Runoff* as the models are trained with data from all catchments across arid to humid groups.

c. *Parameter uncertainty*

In this section, we investigate the effect of (hyper) parameter selection on the stability of our results. This is to address the concern that multiple parameter sets provide equally acceptable model outputs during calibration, while they may yield rather variable predictions (Ebel and Loague 2006), even for such a simple index like a sign of change in mean annual runoff (Melsen et al. 2018). Fowler et al. (2016) particularly addressed this equifinality issue in the framework of DSST and showed that calibration methods often fail to identify parameter sets that are robust over a wide variety of conditions.

All simulations during the evaluation period are repeated with different sets of parameters for the process-based models. In addition to the best parameter values used for the main analyses, we choose the fifth, tenth, and twentieth best-performing parameter sets based on model performance during calibration. As depicted in

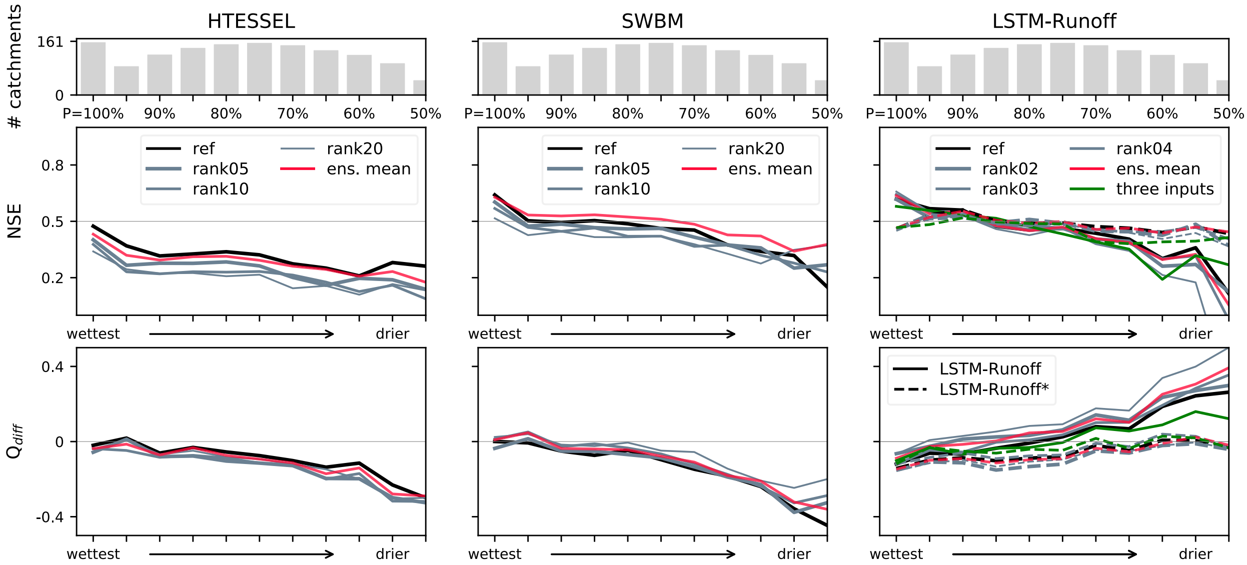

FIG. 5. As in Fig. 2, but for simulations with the fifth, tenth, and twentieth best-performing parameter sets for (left) HTESSEL and (center) SWBM. Black lines (ref) show performance from the best-performing parameter sets, i.e., same as results in Fig. 2. The ensemble mean (red) is obtained from averaged runoff from all four simulations. (right) For LSTM-Runoff, simulations with different model configurations are shown. The model is trained with the second, third, and fourth best-performing hyperparameters, respectively, and their ensemble mean is shown in red. Additionally, LSTM-Runoff is trained with three input variables as for SWBM (green). Note that for simplicity only wet2dry results are shown, while respective dry2wet results can be found in the supplemental material (Fig. S4).

Fig. 5 (left and center panels for HTESSEL and SWBM, respectively), the model results obtained with the different parameter sets are very similar to those with the best (reference) parameter set. This suggests that the calibration methodology overall does not impact our conclusions. In addition, we generate ensemble simulations, i.e., the four simulated runoff outputs from the same model are averaged to obtain a single representative result. As depicted in the relevant figure (red lines), ensemble means show performance comparable with the default simulation, or even better performance as in the case of SWBM. By combining different parameter sets, their individual weaknesses can be compensated. This highlights how multiparameter ensembles might be a potential avenue to increase models' robustness in the future (Yokohata et al. 2012; Orth et al. 2016; Her et al. 2019).

Similarly, LSTM-Runoff models are rerun with different sets of hyperparameters that show comparable performance during model training (Fig. 5, right panel); the second, third, and fourth best-performing hyperparameter sets. The selected ranks are different from those for the process-based models given the fewer number of hyperparameter sets. We additionally train the model with the three input variables same as those used for SWBM. As seen from the process-based models, the overall results remain the same regardless of the selected hyperparameters or input information. Interestingly, we observe that simulations with the three inputs exhibit more reliable performance along the changing conditions. This indicates that using less numbers of forcing variables can prevent LSTM-Runoff from overfitting by permitting it to focus on the dominant processes. As a result, we conclude that our results are not significantly affected by the selection of (hyper)parameters for all three models.

5. Discussion and conclusions

This study assesses the robustness of three different LSMs under changing climatic conditions. The employed models represent physically based, conceptual, and empirical schemes, respectively, and therefore include different levels of knowledge about land surface states and processes, and hence complexity. All three models feature a gradual decrease in overall performance as conditions deviate from the reference, but at different rates depending on their complexity. Our main findings are summarized in Fig. 6. The figure is derived by applying a linear regression to the main result of Fig. 2 and rates (slope) of deterioration of model performance are used as an indication of models' robustness. This analysis highlights our main conclusion, namely, when comparing physically based and conceptual models, the robustness of model performance in transient climate conditions is increased the more physics (e.g., water and energy

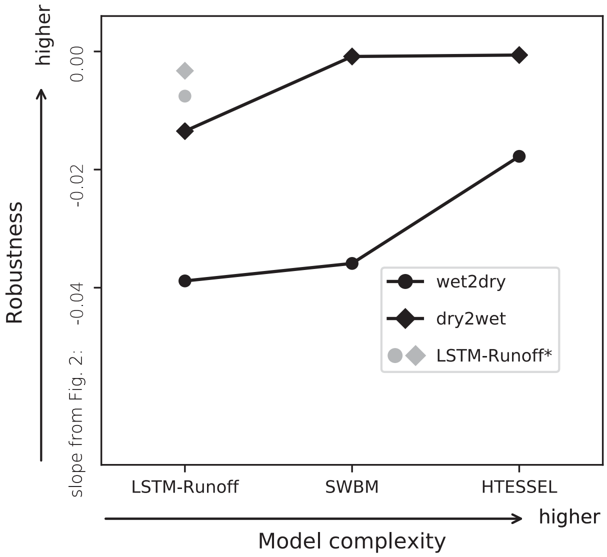

FIG. 6. Relation between model complexity and performance robustness under changing conditions. Robustness is depicted by its inverse relation with slope values of model performance deterioration, which are obtained by applying a linear regression to results of Fig. 2. Note that this figure summarizes the results found from our study; it does not show absolute performance of model schemes.

conservation in HTESSEL) is applied to constrain a model. More importantly, we further find that deteriorated model performance under transient conditions is partially attributable to temporal shifts between hydroclimatic regimes that cannot be adequately described by time-invariant model parameters. LSTM-Runoff can overcome this issue by employing training data from multiple catchments across contrasting conditions, without increasing the amount of training data.

In the light of the relatively short 1-yr periods used for calibration, our study should be regarded as a first-order experiment to trigger further future research. While it is uncommon to calibrate models over such short periods, this allows us to assess model performance under a wide range of climate conditions as the difference between driest and wettest 1-yr periods exceeds that of longer periods. Somewhat addressing the short calibration time period, we run the process-based models several times for any given year. The resulting multiyear simulations enable the models to reach equilibrium states. For LSTM-Runoff, we use training data from all available catchments. Further, recomputing our analysis using 2-yr periods (Fig. S2) illustrates that our conclusions are not significantly impacted by the choice of the length of the time period.

Next to the findings on model robustness, our study constitutes a pioneer effort in comparing process-based and data-driven models. This is inherently difficult as process-based models are developed with knowledge of and from all climatic conditions, and then calibrated for constrained extreme conditions in our study. This is not reproducible for data-driven models as their knowledge is developed purely from the given training conditions during calibration. We address this issue by 1) giving the combined calibration data from all catchments to the data-driven model (LSTM-Runoff) and by 2) giving the combined calibration data, but across all available conditions, from all catchments to the data-driven model (LSTM-Runoff*). The results of LSTM-Runoff show that models of this type suffer clearly from limited performance outside the training conditions, which can be understood from their missing physics foundations. However, the strong robustness of LSTM-Runoff* demonstrates the potential of such machine-learning approaches for modeling under changing conditions. Essentially, they can efficiently "trade space for time." This involves assuming an analogy between extrapolation in time (predictions under climate change) and extrapolation in space (predictions in ungauged basins); i.e., training models with multiple catchments such that models can experience various conditions (e.g., Peel and Blöschl 2011; Singh et al. 2011).

Despite the first-order nature of the analysis, our results have important implications for hydrologic prediction and climate impact studies. While land surface and hydrological models are a primary tool for climate impact assessment in hydrologic systems, predictive errors of the models can be substantial because of a transient climate. This will be particularly the case in regions which are expected to experience significant changes in their hydroclimatic conditions (Berg et al. 2016; Lin et al. 2018). Further, our findings suggest potential difficulties of LSM predictions in semiarid regions, where a regime change between energy- and water-limited conditions is expected. Moreover, these regions are particularly vulnerable to climate change in terms of water availability, e.g., for agricultural, industrial, and domestic demands (Ragab and Prudhomme 2002; Herrera-Pantoja and Hiscock 2015).

Most importantly, the relatively robust performance of the most complex model, HTESSEL, highlights the need for further improved and expanded processes representations in current LSMs. This will also improve the efficacy of LSMs within coupled Earth system models used for climate change projections. The consequent increase of the number of model parameters potentially raises the risk of overparameterization. However, given our findings, this is not a problem as long as the additional parameters introduce stronger physical constraints to the model. In contrast, adding parameters without enhancing the physical constraints,

TABLE A1. Model parameters and their perturbed range for HTESSEL.

Parameter	Description	Range of multiplicative perturbation
Minimum stomatal resistance	Scales leaf area index in the computation of canopy resistance	[0.25, 4]
Soil moisture stress	Determines the shape (e.g., 1 for linear) of dependency of canopy resistance on soil moisture	[0.25, 4]
Total soil depth	Lower boundaries of the particular soil layers; top layer not impacted by perturbations to avoid impacts on the fast thermal response	[0.5, 2]
Van Genuchten alpha	Soil-dependent soil texture parameter in calculation of hydraulic conductivity	[0.25, 4]
Saturated hydraulic conductivity	Soil-dependent parameter; governs vertical percolation of water within the soil profile	[0.25, 4]

as carried out in many conceptual models, does not necessarily lead to improved robustness toward transient conditions (e.g., Vaze et al. 2010; Coron et al. 2012), probably owing to overfitting problems.

For empirical and machine learning-based models, establishing a physics foundation can be realized via a hybrid approach. Such methodology aims to bridge the gap between process-based and data-driven models, by developing algorithms to ensure optimal combinations of respective models, for more secure application in climate research (Reichstein et al. 2019). For instance, the performance metric used to train machine learning models can be modified to account for the physical consistency of the model predictions, i.e., physics-guided machine learning (e.g., Karpatne et al. 2017; Yang et al. 2019). In this way, we might better guarantee the model robustness beyond the climatic conditions over the available data period. In addition, it should be noted that, given the flexibility of LSTM, this study does not aim to provide any conclusive remarks on the extrapolation capacity or absolute performance of LSTM or LSTM-based models. In this context, there are ample possibilities for LSTM-based models to learn a wide range of physics relevant to hydrologic processes (i.e., increasing model complexity). For instance, Kratzert et al. (2019) showed that LSTM can extract information benefiting its runoff modeling performance from static catchment attributes. Therefore, LSTM-based models have great potential to improve their robustness by extracting "hidden" information from diverse data possibly including those that are generally not used in process-based models, e.g., time series of ecological-status data from remote sensing.

While our study indicates benefits of LSM complexity for more robust performance in transient climate, this inspires follow-up questions: how much more complexity will still be useful given increasingly difficult parameter estimation? Which are the key physical mechanisms that are particularly relevant for more reliable future projections? These comprehensive questions, and the potential that correspondingly improved understanding holds, call for a collaborative effort among climate modeling communities, e.g., PILPS or PLUMBER-like projects for model intercomparison and benchmarking under a common framework. As expressed by Best et al. (2015), such project permits "to target areas requiring improvements common to all groups, as well as areas specific to individual modelling groups." Multimodel evaluation in the context of climate change has been attempted by Thirel et al. (2015b), but to a limited extent. Our results emphasize the need to include broader ranges of models within a systematic assessment for comparison of model behaviors and for identification of required degrees of realism. Additionally, model evaluation should be extended to address more variables beyond runoff, such as soil moisture and ET, and furthermore to assess the physical consistency of their simulated interplay. Importantly, more research is needed on key climatic indices that can explain model performance deterioration under transient climate conditions. Indices such as aridity identified in our study, have great potential to guide future model development which enables more reliable climate predictions,

TABLE A2. Model parameters and their prior range for SWBM.

Parameter	Description	Prior range
Water holding capacity	Maximum water storage	[50, 1500]
Runoff function exponent	Sensitivity of normalized runoff to soil moisture	[0.4, 15]
ET function exponent	Sensitivity of ET to soil moisture	[0.03, 1.25]
Maximum ET ratio	Maximum fraction of ET	[0.30, 0.99]
Melting parameter	Speed of snow melting	[0.15, 12]
Runoff delay	Conversion of runoff to streamflow	[0.05, 1.5]

TABLE A3. Considered hyperparameters and their values for LSTM-Runoff.

Hyperparameter	Description	Considered value
Hidden unit	The number of basic computation units (i.e., neurons) per layer	10, 20, 30
Hidden layer	The collection of neurons	1, 2
Look-back	The number of prior time steps (days) of input to predict output at a certain time step	30, 40, 60
Dropout rate	The fraction of units randomly dropped out during training	0.1, 0.2, 0.4
Activation function	Transfer function to convert input signals of a neuron to output signals	Tanh for the hidden layers and rectified linear unit for the last dense layer

however, additional experiments with various models and regions are required to generalize the findings.

Acknowledgments. We gratefully acknowledge the support by the German Research Foundation: Emmy Noether Grant 391059971. This study is a contribution to the Euro-FRIEND project. We thank Gianpaolo Balsamo (ECMWF) for providing the HTESSEL model. We also thank Markus Reichstein and Basil Kraft (Max Planck Institute for Biogeochemistry) for valuable discussion on data-driven modeling. We acknowledge streamflow data from the European water archive in cooperation with the EU-FP6 project WATCH (http://www.eu-watch.org, accessed on 15 May 2020) and the E-OBS dataset established by the EU-FP6 project UERRA (http://www.uerra.eu, accessed on 15 May 2020) and the data providers in the ECA&D project (http://www.ecad.eu, accessed on 15 May 2020). WFDEI data are obtained from the eartH2Observe Water Cycle Integrator (https://wci.earth2observe.eu, accessed on 15 May 2020).

APPENDIX

Model Calibration and Training

a. HTESSEL

We choose five parameters exerting the most influence on model runoff outputs based on the results from sensitivity analysis conducted by Orth et al. (2016) and MacLeod et al. (2016). See Table A1. Note that only a small fraction of the model parameters is selected for better calibration efficiency. Following the previous studies, default values of the parameters are perturbed at once using multiplicative factors between 0.25 and 4 (between 0.5 and 2 for the soil depth). In total 500 sets of multiplicative factors are randomly selected using Latin hypercube sampling (LHS; McKay et al. 1979), which allows to explore the parameter space as completely as possible. The best-performing set of parameter perturbation factors are determined by root-mean-square error (RMSE) between observed and simulated runoff.

b. SWBM

The same calibration strategy as adopted for HTESSEL is applied to SWBM. All six parameters of the model are optimized. Ranges of parameter spaces are specified based on the range of calibrated values found in Orth and Seneviratne (2013) and Orth and Seneviratne (2015). See also Table A2. In total 500 sets of parameters are randomly sampled using LHS, and RMSE is used to verify model performance during calibration, as done for HTESSEL.

c. LSTM-Runoff

Following a calibration strategy for the machine learning, known as tuning of hyperparameters (Chollet 2017), training data during the reference period, i.e., 1 year × 161 catchments are divided into two splits to train and validate the LSTM-based model. Training-validation is a task to find an optimal set of hyperparameters during model training through cross validation on a portion of training data. In this study, hyperparameters are selected through a grid search with k-fold cross validation ($k = 10$); each fold is held out to evaluate performance of the model trained on the remaining $K - 1$ folds. The final model configuration is decided by choosing the hyperparameter set that resulted in the smallest average RMSE across k-folds among all considered configurations. The hyperparameter values considered in this study are listed in Table A3.

REFERENCES

Abramowitz, G., R. Leuning, M. Clark, and A. Pitman, 2008: Evaluating the performance of land surface models. *J. Climate*, **21**, 5468–5481, https://doi.org/10.1175/2008JCLI2378.1.

Balsamo, G., A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, and A. K. Betts, 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *J. Hydrometeor.*, **10**, 623–643, https://doi.org/10.1175/2008JHM1068.1.

——, and Coauthors, 2015: ERA-Interim/Land: A global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.*, **19**, 389–407, https://doi.org/10.5194/hess-19-389-2015.

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

Berg, A., and Coauthors, 2016: Land-atmosphere feedbacks amplify aridity increase over land under global warming. *Nat. Climate Change*, **6**, 869–874, https://doi.org/10.1038/nclimate3029.

Best, M. J., and Coauthors, 2015: The plumbing of land surface models: Benchmarking model performance. *J. Hydrometeor.*, **16**, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1.

Beven, K., 1989: Changing ideas in hydrology—The case of physically-based models. *J. Hydrol.*, **105**, 157–172, https://doi.org/10.1016/0022-1694(89)90101-7.

Boone, A., and Coauthors, 2009: The AMMA Land Surface Model Intercomparison Project (ALMIP). *Bull. Amer. Meteor. Soc.*, **90**, 1865–1880, https://doi.org/10.1175/2009BAMS2786.1.

Brigode, P., L. Oudin, and C. Perrin, 2013: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change? *J. Hydrol.*, **476**, 410–425, https://doi.org/10.1016/j.jhydrol.2012.11.012.

Broderick, C., T. Matthews, R. L. Wilby, S. Bastola, and C. Murphy, 2016: Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resour. Res.*, **52**, 8343–8373, https://doi.org/10.1002/2016WR018850.

Budyko, M., 1974: *Climate and Life*. Academic Press, 507 pp.

Chollet, F., 2017: *Deep Learning with Python*. 1st ed. Manning Publications Co., 384 pp.

Cornes, R. C., G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones, 2018: An ensemble version of the e-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.*, **123**, 9391–9409, https://doi.org/10.1029/2017JD028200.

Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx, 2012: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resour. Res.*, **48**, W05552, https://doi.org/10.1029/2011WR011721.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Denissen, J. M., A. J. Teuling, M. Reichstein, and R. Orth, 2020: Critical soil moisture derived from satellite observations over Europe. *J. Geophys. Res. Atmos.*, e2019JD031672, https://doi.org/10.1029/2019JD031672.

Dirmeyer, P. A., 2011: A history and review of the Global Soil Wetness Project (GSWP). *J. Hydrometeor.*, **12**, 729–749, https://doi.org/10.1175/JHM-D-10-05010.1.

——, A. J. Dolman, and N. Sato, 1999: The pilot phase of the global soil wetness project. *Bull. Amer. Meteor. Soc.*, **80**, 851–878, https://doi.org/10.1175/1520-0477(1999)080<0851:TPPOTG>2.0.CO;2.

Ebel, B. A., and K. Loague, 2006: Physics-based hydrologic-response simulation: Seeing through the fog of equifinality. *Hydrol. Processes*, **20**, 2887–2900, https://doi.org/10.1002/hyp.6388.

ECMWF, 2016: Part IV: Physical processes. ECMWF, IFS Doc. 4, accessed 1 January 2019, 223 pp., https://www.ecmwf.int/node/17117.

Fowler, K. J. A., M. C. Peel, A. W. Western, L. Zhang, and T. J. Peterson, 2016: Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Resour. Res.*, **52**, 1820–1846, https://doi.org/10.1002/2015WR018068.

Glorot, X., and Y. Bengio, 2010: Understanding the difficulty of training deep feedforward neural networks. *Proc. 13th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy, AISTATS, 249–246, http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf.

Haddeland, I., and Coauthors, 2011: Multimodel estimate of the global terrestrial water balance: Setup and first results. *J. Hydrometeor.*, **12**, 869–884, https://doi.org/10.1175/2011JHM1324.1.

Henderson-Sellers, A., Z.-L. Yang, and R. E. Dickinson, 1993: The Project for Intercomparison of Land-Surface Parameterization Schemes. *Bull. Amer. Meteor. Soc.*, **74**, 1335–1349, https://doi.org/10.1175/1520-0477(1993)074<1335:TPFIOL>2.0.CO;2.

——, A. J. Pitman, P. K. Love, P. Irannejad, and T. H. Chen, 1995: The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3. *Bull. Amer. Meteor. Soc.*, **76**, 489–503, https://doi.org/10.1175/1520-0477(1995)076<0489:TPFIOL>2.0.CO;2.

Her, Y., S.-H. Yoo, J. Cho, S. Hwang, J. Jeong, and C. Seong, 2019: Uncertainty in hydrological analysis of climate change: Multiparameter vs. multi-GCM ensemble predictions. *Sci. Rep.*, **9**, 4974, https://doi.org/10.1038/s41598-019-41334-7.

Herrera-Pantoja, M., and K. Hiscock, 2015: Projected impacts of climate change on water availability indicators in a semi-arid region of central Mexico. *Environ. Sci. Policy*, **54**, 81–89, https://doi.org/10.1016/j.envsci.2015.06.020.

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

Hofstra, N., M. Haylock, M. New, and P. D. Jones, 2009: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. *J. Geophys. Res.*, **114**, D21101, https://doi.org/10.1029/2009JD011799.

Hu, C., Q. Wu, H. Li, S. Jian, N. Li, and Z. Lou, 2018: Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water*, **10**, 1543, https://doi.org/10.3390/w10111543.

Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55.

Jiménez, C., and Coauthors, 2011: Global intercomparison of 12 land surface heat flux estimates. *J. Geophys. Res.*, **116**, D02102, https://doi.org/10.1029/2010JD014545.

Karpatne, A., W. Watkins, J. Read, and V. Kumar, 2017: Physics-Guided Neural Networks (PGNN): An application in lake temperature modeling. arXiv, 11 pp., https://arxiv.org/abs/1710.11431.

Klemeš, V., 1986: Operational testing of hydrological simulation models. *Hydrol. Sci. J.*, **31**, 13–24, https://doi.org/10.1080/02626668609491024.

Kling, H., P. Stanzel, M. Fuchs, and H.-P. Nachtnebel, 2015: Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates. *Hydrol. Sci. J.*, **60**, 1374–1393, https://doi.org/10.1080/02626667.2014.959956.

Koster, R. D., and P. C. D. Milly, 1997: The interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models. *J. Climate*, **10**, 1578–1591, https://doi.org/10.1175/1520-0442(1997)010<1578:TIBTAR>2.0.CO;2.

——, and S. P. Mahanama, 2012: Land surface controls on hydroclimatic means and variability. *J. Hydrometeor.*, **13**, 1604–1620, https://doi.org/10.1175/JHM-D-12-050.1.

——, and Coauthors, 2006: GLACE: The Global Land–Atmosphere Coupling Experiment. Part I: Overview. *J. Hydrometeor.*, **7**, 590–610, https://doi.org/10.1175/JHM510.1.

Kratzert, F., D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, 2018: Rainfall–runoff modelling using Long Short-Term

Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.*, **22**, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018.

——, ——, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, 2019: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.*, **23**, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019.

LeCun, Y. A., L. Bottou, G. B. Orr, and K.-R. Müller, 2012: Efficient BackProp. *Neural Networks: Tricks of the Trade*, 2nd ed. Springer, 9–48, https://doi.org/10.1007/978-3-642-35289-8_3.

Li, C. Z., L. Zhang, H. Wang, Y. Q. Zhang, F. L. Yu, and D. H. Yan, 2012: The transferability of hydrological models under nonstationary climatic conditions. *Hydrol. Earth Syst. Sci.*, **16**, 1239–1254, https://doi.org/10.5194/hess-16-1239-2012.

Li, H., S. Beldring, and C.-Y. Xu, 2015: Stability of model performance and parameter values on two catchments facing changes in climatic conditions. *Hydrol. Sci. J.*, **60**, 1317–1330, https://doi.org/10.1080/02626667.2014.978333.

Lin, L., A. Gettelman, Q. Fu, and Y. Xu, 2018: Simulated differences in 21st century aridity due to different scenarios of greenhouse gases and aerosols. *Climatic Change*, **146**, 407–422, https://doi.org/10.1007/s10584-016-1615-3.

MacLeod, D. A., H. L. Cloke, F. Pappenberger, and A. Weisheimer, 2016: Improved seasonal prediction of the hot summer of 2003 over Europe through better representation of uncertainty in the land surface. *Quart. J. Roy. Meteor. Soc.*, **142**, 79–90, https://doi.org/10.1002/qj.2631.

Materia, S., P. A. Dirmeyer, Z. Guo, A. Alessandri, and A. Navarra, 2010: The sensitivity of simulated river discharge to land surface representation and meteorological forcings. *J. Hydrometeor.*, **11**, 334–351, https://doi.org/10.1175/2009JHM1162.1.

McKay, M. D., R. J. Beckman, and W. J. Conover, 1979: Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245, https://doi.org/10.1080/00401706.1979.10489755.

Melsen, L. A., N. Addor, N. Mizukami, A. J. Newman, P. J. J. F. Torfs, M. P. Clark, R. Uijlenhoet, and A. J. Teuling, 2018: Mapping (dis) agreement in hydrologic projections. *Hydrol. Earth Syst. Sci.*, **22**, 1775–1791, https://doi.org/10.5194/hess-22-1775-2018.

Merz, R., J. Parajka, and G. Blöschl, 2011: Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resour. Res.*, **47**, W02531, https://doi.org/10.1029/2010WR009505.

Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer, 2008: Stationarity is dead: Whither water management? *Science*, **319**, 573–574, https://doi.org/10.1126/science.1151915.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE*, **50**, 885–900, https://doi.org/10.13031/2013.23153.

Motovilov, Y. G., L. Gottschalk, K. Engeland, and A. Rodhe, 1999: Validation of a distributed hydrological model against spatial observations. *Agric. For. Meteor.*, **98–99**, 257–277, https://doi.org/10.1016/S0168-1923(99)00102-1.

Nash, J., and J. Sutcliffe, 1970: River flow forecasting through conceptual models Part I—A discussion of principles. *J. Hydrol.*, **10**, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6.

Orth, R., and S. I. Seneviratne, 2013: Propagation of soil moisture memory to streamflow and evapotranspiration in Europe. *Hydrol. Earth Syst. Sci.*, **17**, 3895–3911, https://doi.org/10.5194/hess-17-3895-2013.

——, and ——, 2015: Introduction of a simple-model-based land surface dataset for Europe. *Environ. Res. Lett.*, **10**, 044012, https://doi.org/10.1088/1748-9326/10/4/044012.

——, and G. Destouni, 2018: Drought reduces blue-water fluxes more strongly than green-water fluxes in Europe. *Nat. Commun.*, **9**, 3602, https://doi.org/10.1038/s41467-018-06013-7.

——, M. Staudinger, S. I. Seneviratne, J. Seibert, and M. Zappa, 2015: Does model performance improve with complexity? A case study with three hydrological models. *J. Hydrol.*, **523**, 147–159, https://doi.org/10.1016/j.jhydrol.2015.01.044.

——, E. Dutra, and F. Pappenberger, 2016: Improving weather predictability by including land surface model parameter uncertainty. *Mon. Wea. Rev.*, **144**, 1551–1569, https://doi.org/10.1175/MWR-D-15-0283.1.

Peel, M. C., and G. Blöschl, 2011: Hydrological modelling in a changing world. *Prog. Phys. Geogr.*, **35**, 249–261, https://doi.org/10.1177/0309133311402550.

Perrin, C., C. Michel, and V. Andréassian, 2001: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol.*, **242**, 275–301, https://doi.org/10.1016/S0022-1694(00)00393-0.

Pilgrim, D. H., T. G. Chapman, and D. G. Doran, 1988: Problems of rainfall-runoff modelling in arid and semiarid regions. *Hydrol. Sci. J.*, **33**, 379–400, https://doi.org/10.1080/02626668809491261.

Ragab, R., and C. Prudhomme, 2002: SW—Soil and water: Climate change and water resources management in arid and semi-arid regions: Prospective and challenges for the 21st century. *Biosyst. Eng.*, **81**, 3–34, https://doi.org/10.1006/bioe.2001.0013.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, and DMIP Participants, 2004: Overall distributed model intercomparison project results. *J. Hydrol.*, **298**, 27–60, https://doi.org/10.1016/j.jhydrol.2004.03.031.

Refsgaard, J. C., and Coauthors, 2014: A framework for testing the ability of models to project climate change and its impacts. *Climatic Change*, **122**, 271–282, https://doi.org/10.1007/s10584-013-0990-2.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Sahoo, B. B., R. Jha, A. Singh, and D. Kumar, 2019: Long Short-Term Memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophys.*, **67**, 1471–1481, https://doi.org/10.1007/s11600-019-00330-1.

Seibert, J., 2003: Reliability of model predictions outside calibration conditions. *Hydrol. Res.*, **34**, 477–492, https://doi.org/10.2166/nh.2003.0019.

Seiller, G., F. Anctil, and C. Perrin, 2012: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrol. Earth Syst. Sci.*, **16**, 1171–1189, https://doi.org/10.5194/hess-16-1171-2012.

——, I. Hajji, and F. Anctil, 2015: Improving the temporal transposability of lumped hydrological models on twenty diversified U.S. Watersheds. *J. Hydrol. Reg. Stud.*, **3**, 379–399, https://doi.org/10.1016/j.ejrh.2015.02.012.

Seneviratne, S. I., and Coauthors, 2013: Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. *Geophys. Res. Lett.*, **40**, 5212–5217, https://doi.org/10.1002/grl.50956.

Singh, R., T. Wagener, K. van Werkhoven, M. E. Mann, and R. Crane, 2011: A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate - accounting for changing watershed behavior. *Hydrol. Earth Syst. Sci.*, **15**, 3591–3603, https://doi.org/10.5194/hess-15-3591-2011.

Stahl, K., and Coauthors, 2010: Streamflow trends in Europe: Evidence from a dataset of near-natural catchments. *Hydrol. Earth Syst. Sci.*, **14**, 2367–2382, https://doi.org/10.5194/hess-14-2367-2010.

Tegegne, G., D. K. Park, and Y.-O. Kim, 2017: Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *J. Hydrol. Reg. Stud.*, **14**, 49–66, https://doi.org/10.1016/j.ejrh.2017.10.002.

Thirel, G., V. Andréassian, and C. Perrin, 2015a: On the need to test hydrological models under changing conditions. *Hydrol. Sci. J.*, **60**, 1165–1173, https://doi.org/10.1080/02626667.2015.1050027.

——, and Coauthors, 2015b: Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrol. Sci. J.*, **60**, 1184–1199, https://doi.org/10.1080/02626667.2014.967248.

Troch, P. A., and Coauthors, 2009: Climate and vegetation water use efficiency at catchment scales. *Hydrol. Processes*, **23**, 2409–2414, https://doi.org/10.1002/hyp.7358.

Vaze, J., D. Post, F. Chiew, J.-M. Perraud, N. Viney, and J. Teng, 2010: Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies. *J. Hydrol.*, **394**, 447–457, https://doi.org/10.1016/j.jhydrol.2010.09.018.

Vormoor, K., M. Heistermann, A. Bronstert, and D. Lawrence, 2018: Hydrological model parameter (in)stability – ‘‘Crash testing’’ the HBV model under contrasting flood seasonality conditions. *Hydrol. Sci. J.*, **63**, 991–1007, https://doi.org/10.1080/02626667.2018.1466056.

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

Wei, J., and P. A. Dirmeyer, 2010: Toward understanding the large-scale land-atmosphere coupling in the models: Roles of different processes. *Geophys. Res. Lett.*, **37**, L19707, https://doi.org/10.1029/2010GL044769.

Wilby, R. L., 2005: Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrol. Processes*, **19**, 3201–3219, https://doi.org/10.1002/hyp.5819.

Xu, C.-y., E. Widén, and S. Halldin, 2005: Modelling hydrological consequences of climate change—Progress and challenges. *Adv. Atmos. Sci.*, **22**, 789–797, https://doi.org/10.1007/BF02918679.

Yang, T., F. Sun, P. Gentine, W. Liu, H. Wang, J. Yin, M. Du, and C. Liu, 2019: Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environ. Res. Lett.*, **14**, 114027, https://doi.org/10.1088/1748-9326/ab4d5e.

Yokohata, T., J. D. Annan, M. Collins, C. S. Jackson, M. Tobis, M. J. Webb, and J. C. Hargreaves, 2012: Reliability of multi-model and structurally different single-model ensembles. *Climate Dyn.*, **39**, 599–616, https://doi.org/10.1007/s00382-011-1203-1.

Zhang, D., J. Lin, Q. Peng, D. Wang, T. Yang, S. Sorooshian, X. Liu, and J. Zhuang, 2018: Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *J. Hydrol.*, **565**, 720–736, https://doi.org/10.1016/j.jhydrol.2018.08.050.