Can hydrological models benefit from using global soil moisture, evapotranspiration, and runoff products as calibration targets?

Yiwen Mei[1,*], Juliane Mai[2], Hong Xuan Do[1,3], Andrew Gronewold[1,4], Howard Reeves[5], Sandra Eberts[6], Richard Niswonger[7], R. Steven Regan[8], and Randall J. Hunt[9]

[1] School for Environment and Sustainability, University of Michigan, Ann Arbor, MI, USA

[2] Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada

[3] Faculty of Environment and Natural Resources, Nong Lam University, Ho Chi Minh City, Vietnam

[4] Department of Earth and Environmental Science, University of Michigan, Ann Arbor, MI, USA

[5] U.S. Geological Survey, Upper Midwest Water Science Center, Lansing, MI, USA

[6] U.S. Geological Survey, Earth Systems Processes Division, Water Resources Mission Area, Columbus, OH, USA

[7] U.S. Geological Survey, Integrated Modeling and Prediction Division, Water Resources Mission Area, Menlo Park, CA, USA

[8] U.S. Geological Survey, Integrated Modeling and Prediction Division, Water Resources Mission Area, Lakewood, CO, USA

[9] U.S. Geological Survey, Upper Midwest Water Science Center, Madison, WI, USA

Corresponding author: Yiwen Mei (yiwenm@umich.edu; yiwen.mei@uconn.edu)

Key Points:

- Using soil moisture in addition to streamflow to constrain hydrological model calibration improves the evapotranspiration simulation.
- The global gridded runoff products show higher potential in streamflow calibration for larger catchments.
- Ternary diagram is used to visualize performances of three hydrological variables considering all possible variable importance.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2022WR032064.

Abstract

Hydrological models are usually calibrated to in-situ streamflow observations with reasonably long and uninterrupted records. This is challenging for poorly gage or ungaged basins where such information is not available. Even for gaged basins, the single-objective calibration to gaged streamflow cannot guarantee reliable forecasts because, as has been documented elsewhere, the inverse problem is mathematically ill-posed. Therefore, inclusion of other observations, and reproduction of other hydrological variables beyond streamflow, become critical components of accurate hydrological forecasting. In this study, six single- and multi-objective model calibration schemes based on different combinations of gaged streamflow, global-scale gridded soil moisture, actual evapotranspiration (ET), and runoff products are used for the calibration of a process-based hydrological model for 20 catchments located within the Lake Michigan watershed, of the Laurentian Great Lakes. Results show that the addition of gridded soil moisture to gaged streamflow in model calibration improves the ET simulation performance for most of the catchments, leading to the overall best performing models. The monthly streamflow simulation performance for the experiments using gridded runoff products to inform the model is outperformed by those using the gaged streamflow, but the discrepancy is mitigating with increasing catchment scale. A new visualization method that effectively synthesizes model performance for the simulations of streamflow, soil moisture, and ET was also proposed. Based on the method, it is revealed that the streamflow simulation performance is relatively weak for baseflow-dominated catchments; overall, the 20 catchment models simulate streamflow and ET better than soil moisture.

1. Introduction

Hydrological modeling is the primary approach for estimating continuous terrestrial hydrological components at different spatial and temporal scales. Models can be used to simulate historical changes in regional water cycles (Regan et al., 2019), provide necessary input data for other earth system science investigations – e.g., environmental flow (Liu et al., 2016), drought analysis (Van Loon & Laaha 2015)), perform scenario-based predictions of the future (Champagne et al., 2020), or inform water resources management practices and decision making (He et al., 2021). Watershed flow comprises many processes occurring at a range of scales, requiring realistic hydrological models that contain many parameters that cannot be measured in practice. These parameters are usually attained by calibrating to in-situ streamflow that provides an integrated estimate of all hydrological partitioning upstream of a gage. Yet the hydrological partitioning at the gage cannot be uniquely identified and is further complicated by uncertainty in other important watershed hydrological components such as evapotranspiration (ET), soil moisture, and recharge (Dembélé et al., 2020; Bai et al., 2018; Rakovec et al., 2016). Such uncertainty can lead to unrealistic model parameters that compensate for parameter simplification errors in the model and biases in other hydrological processes such as climate forcing or incorrect representation of hydrological processes in the model (Knoben et al., 2020; McMillan et al., 2018; Beven, 2006). In addition, the streamflow-only calibration approach relies on one type of calibration target, which precludes its application in regions where gaged streamflow data are not available. These so-called ungaged catchments comprise most of the land area of the globe (Do et al., 2018). The integration of remote sensing-based observations of soil moisture, ET, and streamflow in model calibration is a potential approach to provide a comprehensive calibration scheme that considers many components of the hydrological cycles relative to a streamflow-only calibration approach.

Previous studies have investigated the added value of using remote sensing-based hydrological observations in addition to gaged streamflow in model calibration (see **Table 1** for references). **Table 1** summarizes a wide range of such studies and highlights the model performance of different calibration schemes that include the use of remote sensing-based soil moisture, ET, runoff, and terrestrial water storage (TWS) compared to streamflow-only calibration. **Table 1** lists 16 representative studies published within the last 10 years that used multiple datasets to inform hydrological models. Based on the table, some common findings and gaps can be identified (note that the results might be model-/region-specific as the studies are done with different models and in different regions). The first common finding is that the use of additional remote sensing-based variables besides gaged streamflow does not improve streamflow prediction (see column "Q/R" and rows "Streamflow + additional variables"); however, these studies show an improved performance for the additional variables used for calibration (see columns "ET", "S", and "TWS" and rows "Streamflow + additional variables"). Analogously, the sole use of remote sensing-based gridded ET, soil moisture, or/and TWS in calibration generally does not improve the prediction of streamflow but does improve the simulations of the respective variables (see rows "Remote sensing-only"). Another finding worth mentioning is that four studies (Dembélé et al., 2020; Bai et al., 2018; Rakovec et al., 2016; Kunnath-Poovakka et al., 2016) demonstrate that adding TWS improves prediction of ET (see rows "Streamflow+TWS" and "Streamflow+S,TWS"). We will call this phenomenon "cross-benefit" hereafter, referring to a variable benefitting from the addition of another variable in model calibration, considering a baseline model using a streamflow-only calibration.

Three gaps are identified in previous studies on this topic (**Table 1**). The first gap is related to the "cross-benefit" phenomenon introduced in the previous paragraph. With regards to studies that have used "Streamflow + additional variables" to calibrate models, these studies did not evaluate if the inclusion of soil moisture can improve the model's prediction of ET or vice versa. Second, the potential of using global-/regional-scale gridded runoff products (GRPs) in model calibration is rarely investigated (see row "ET,R" under "Gridded product only" in **Table 1**). GRPs refer to the observation-based runoff products produced from downscaling gaged streamflow using statistical or machine learning-based techniques (Ghiggi et al., 2019; Hobeichi et al., 2019), not model simulations. The use of GRPs would help to train models at ungaged locations given its space-time continuity. To our knowledge, Xie et al. (2021) is the only study that used a GRP to calibrate a hydrological model. Xie et al. (2021) shows that the combined use of a GRP and an ET product improved the prediction of ET. However, this approach made the streamflow prediction worse. Two questions arise from this: how well does a model perform in terms of soil moisture simulation when it is trained with gridded runoff instead of gaged streamflow? Which global GRP leads to more reasonable overall model performance? Third, although multiple hydrological variables are considered for calibration, few studies investigate the overall modeling performance that can be used to represent for all their targeting variables. Among those which come up with a synthetic performance metric, only equal weighting factors for the variable's objective functions are considered; see for example Herman et al. (2018) and Kunnath-Poovakka et al. (2016) among others. A research question arising from this observation is if and to what extent does the overall model performance change when different weights are assigned to the performance of different variables? This is not easy to answer as there is no effective method to show all this information in a concise manner (e.g., one single plot).

Table 1. Representative studies within the last 10 years on hydrological model calibration using global-scale gridded runoff (R), soil moisture (S), ET, and TWS products in addition to gaged streamflow (Q). The rows are the different types of calibration schemes grouped into two categories. The first category uses Q plus one or more additional variables, which can be ET, S, and TWS. The "Gridded product-only" category is streamflow-free calibration schemes. The columns are the variables evaluated by the studies. Each study was assigned into one of the three subsets regarding model performance: improve (denoted by "↑"), no obvious change (denoted by "↔"), and decrease (denoted by "↓"). Note that these performance codes reflect the general tendency of results documented by the studies. Bolded citation indicates the use of a multi-objective optimization algorithm for the study. Name code for every representative study is provided in the bottom row (note that the results might be model-/region-specific as the studies are done with different models and in different regions).

		Variable performance is evaluated for											
		Q/R (16 studies)			ET (10 studies)			S (3 studies)			TWS (6 studies)		
		↑	↔	↓	↑	↔	↓	↑	↔	↓	↑	↔	↓
Variable(s) model is calibrated with	Streamflow + additional variables (13 studies)												
	+ET	R2018	**D2018** **H2018**	X2021 H2020 R2013 **LL2012**	X2021 **D2018** **H2018** R2018 R2013	**LL2012**							**LL2012**
	+S	L2018 R2016						R2016					
	+TWS		**B2018** **Y2017** Ra2016	**LL2012**	**B2018** Ra2016	**LL2012**					**B2018** **Y2017** Ra2016	**LL2012**	
	+ET,TWS		D2020	H2020 **LL2012**	D2020	**LL2012**		D2020			D2020		**LL2012**
	+S,TWS		D2020		D2020			D2020				D2020	
	+ET,S		D2020		D2020			D2020			D2020		
	+ET,S,TWS		D2020		D2020			D2020			D2020		

Gridded product-only (9 studies)	ET	**D2018**	**H2018** LL2017 KP2016 R2013 **LL2012**	**D2018** **H2018** KP2016 R2013 **LL2012**			KP2016			**LL2012**
	S		LL2017 KP2016	KP2016		KP2016				
	TWS		**M2018** **LL2012**		**LL2012**			**LL2012**	**M2018**	
	ET,S		LL2017 KP2016	KP2016		KP2016				
	ET,TWS		**LL2012**	**LL2012**				**LL2012**		
	ET,S,TWS		D2020	D2020		D2020		D2020		
	ET,R		X2021	X2021						

Name code of study:

X2021 – Xie et al. (2021)
D2020 – Dembélé et al. (2020)
H2020 – Huang et al. (2020)
B2018 – Bai et al. (2018)
D2018 – Demirel et al. (2018)
H2018 – Herman et al. (2018)
L2018 – Li et al. (2018)
M2018 – Mostafaie et al. (2018)
R2018 – Rajib et al. (2018)
LL2017 – López López et al. (2017)
Y2017 – Yassin et al. (2017)
KP2016 – Kunnath-Poovakka et al. (2016)
R2016 – Rajib et al. (2016)
Ra2016 – Rakovec et al. (2016)
R2013 – Rientjes et al. (2013)
LL2012 – Livneh & Lettenmaier (2012)

This study investigates knowledge gaps in multi-objective model calibration. Specifically, we performed an inter-comparison of model performance of six multi-objective calibration schemes using different combinations of gaged streamflow, global-scale gridded soil moisture, ET, and runoff products. To highlight the benefit of each calibration scheme, we used the results obtained from streamflow-only calibration as the baseline. Further, the potential of two GRPs in model calibration is investigated. A new method is introduced to visualize the combined performance of three modeled variables with different weights. This new method provides a more intuitive interpretation of the model performance. Three objectives are identified for our study:

A. Which additional gridded variables (soil moisture or/and ET) used in calibration result in the best overall model performance? Do we see any "cross-benefit"?

B. What is the relative performance between models calibrated using different global-scale GRPs and gaged streamflow? and

C. How is model performance varied with different weights applied to the objective functions of the hydrological variable?

2. Study catchments and datasets

2.1. Catchments in Lake Michigan watershed

The Lake Michigan watershed of the Laurentian Great Lakes has a drainage area of 173,683 km^2 of which 33% is Lake Michigan itself (**Figure 1**). For this study, 20 independent catchments (i.e., non-nested catchments) ranging from 90 km^2 (catchment i, Menomonee River at Menomonee Falls) to 15,410 km^2 (catchment g, Fox River at Appleton) are selected and set up as independent instances of a hydrological model (section 3.1). The catchments were selected from the U.S. Geological Survey (USGS) GAGES-II dataset (Falcone, 2017) with relatively low human interference (**Table S1** in supporting information). Only gages without back-water from the lake were selected. The watershed is characterized by a mild topography with the mean elevation ranging from 209 m (basin j, Trail Creek at Michigan City) to 413 m (basin b, Escanaba River at Cornell). The 20 streamflow gages also have less than 10% of missing data within the period of 2000 to 2020.

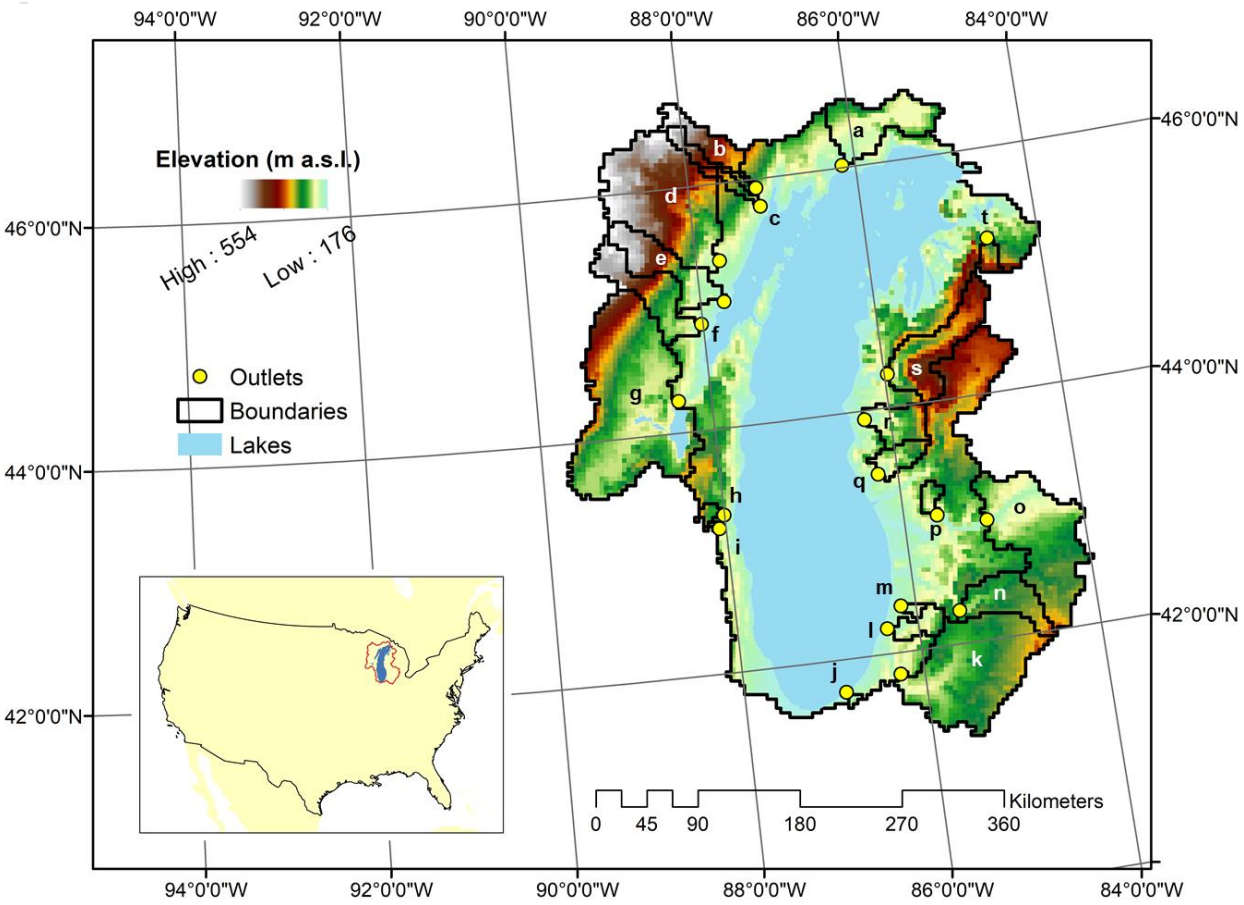


Figure 1. Boundaries and outlets of the 20 catchments (labeled a to t) used in this study overlain on Digital Elevation Model data. Locations of Lake Michigan and the watershed within the Continental United States are outlined in the subpanel.

2.2. Data for hydrological modeling and preprocessing

The modeling exercises of this study require a wide variety of data products (see the product list in **Table S2** in supporting information). The studied catchments are delineated using the hydrographic data (flow direction and accumulation) derived from the Digital Elevation Model (DEM) of HydroSHEDS (Lehner et al., 2008) and the USGS stream gage locations (Falcone, 2017). The 19-class land cover classification used in our study is one of the North America Land Cover Monitoring System (NALCMS) products that is based on the 2010 Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery (Latifovic et al., 2012). The Global Man-made Impervious Surface (GMIS) Dataset from Landsat for 2010 is used to determine the impervious surface ratio (Brown de Colstoun et al., 2017). The MOD44B product from MODIS (DiMiceli et al., 2015) is used to derive the vegetation cover percentage parameters for the summer and winter seasons. The six-layered (0-2m) soil textural profile from SoilGrids (Hengl et al., 2017) is used to derive the soil hydraulic parameters. Forcing data adopted for our hydrological modeling is from Daymet (Thornton et al., 2021). Daymet was developed by interpolating daily meteorological observations in the Global Historical Climatology Network Daily (GHCN-Daily) dataset (Menne et al., 2012) with a truncated Gaussian filter and digital elevation model (Thornton et al., 1997).

Six products are used for model calibration and evaluation. This study uses various calibration schemes, where each scheme uses a different dataset to estimate the runoff generation parameters (section 3.3.1). These calibration datasets are gaged streamflow and gridded soil moisture, ET, and runoff products. The daily streamflow measurements are downloaded from USGS for the 20 study catchments. All station records cover the period from 2000 through 2020.

The SoMo.ml is a recently released observed-based global gridded soil moisture product produced by downscaling soil moisture measurements from more than 1,000 stations across the globe using the long short-term memory neural network (O & Orth, 2021). Dynamical meteorological forcing from the past 365 days and static climate and land surface characteristics are used as predictors. The soil moisture product represents the volumetric water content for three depth intervals (0-10cm, 10-30cm, and 30-50cm). The accuracy of SoMo.ml is shown to be better than two satellite-based and one model-based soil moisture products (O & Orth, 2021).

Another reference product used in model calibration is the Global Land Evaporation Amsterdam Model (GLEAM) by Martens et al. (2017). In GLEAM, actual ET is considered as the sum of bare soil evaporation, transpiration, open-water evaporation, interception loss, and sublimation. Bare soil evaporation or transpiration is adjusted downward from potential evaporation by a stress factor estimated from microwave vegetation optical depth (i.e., water content in vegetation) and root-zone soil moisture calculated from a multilayer water balance algorithm (Miralles et al., 2011). Potential evaporation is calculated based on the Priestley and Taylor (1972) equation with observed surface net radiation and surface air temperature from satellite remote sensing as inputs. Evaporation from open-water is assumed equal to potential evaporation. Interception loss is estimated independently using the refined Gash analytical model (Valente et al., 1997). The estimation of sublimation is based on the snow-water equivalent from the European Space Agency GLOBSNOW product (Luojus et al., 2013). GLEAM ET estimation is shown to have the lowest degree of uncertainty and relatively high accuracy in an evaluation study with other 11 ET products from either modeling or remote sensing over the Continental United States (Xu et al., 2019).

Two global-scale GRPs were also considered in our study. The first one is the Global Runoff Reconstruction (GRUN) dataset (Ghiggi et al., 2019). There are three main procedures to produce GRUN. First, streamflow records from 7,264 catchments, ranging from 10 to 2,500 km^2, are gridded to the cylindrical equal-area (CEA) grid with a 50 km resolution. For every CEA grid, the method uses the median of monthly streamflow from catchments intersecting the CEA grid to represent runoff of the grid cell for the month. These result in runoff records for 5,094 CEA grids, covering 8.5% of the land area. In a second step, 60% of these "observed runoff" estimates are used to train Random Forest (RF) regression models with monthly precipitation and air temperature from the past 6 months as the predictors. The trained RF models are then used to produce monthly runoff for the other CEA grids. The last two steps are repeated 50 times to produce 50 reconstructed runoff to test the sensitivity of the RF model to the training data and to produce the ensemble mean of monthly runoff.

Another global GRP is the Linear Optimal Runoff Aggregate (LORA) dataset (Hobeichi et al., 2019), derived by merging 11 runoff/streamflow estimates from eight Global Hydrological Models (GHM) produced as part of the eartH2Observe project (http://www.earth2observe.eu/). The method starts by constructing linear combinations of the GHM runoff/streamflow that minimizes the mean square difference with the observed streamflow at 596 catchments. These catchments are called donor catchments and their optimal weights are further transferred to other

catchments without observed streamflow data, called receptor catchments. A receptor catchment receives the optimal weights from three donor catchments with the highest similarity – a similarity index defined based on the aridity index, fractions of forest and snow cover, soil clay content, surface slope, and annual averages of precipitation and potential evaporation (Beck et al., 2016). Runoff for a receptor catchment is then estimated by applying the optimal weights to the GHM runoff/streamflow of the catchment.

3. Experimental designs

3.1. Calculations of evapotranspiration, soil moisture, and streamflow in PRMS

The hydrological model used in our study is a process-based distributed deterministic watershed model called Precipitation-Runoff Modeling System (PRMS) version 5.2.0 (Markstrom et al., 2015) within the Ground-water and Surface-water FLOW (GSFLOW) modeling platform (Markstrom et al., 2008). PRMS is designed for the simulation of hydrological processes including evaporation, transpiration, runoff, infiltration, interflow, and groundwater flow as determined by the energy and water budgets of the plant canopy, snowpack, and soil zone based on distributed climate information (Markstrom et al., 2015). The hydrological processes are modeled as a series of reservoirs (plant canopy interception, snowpack, soil zone, impervious zone, subsurface, and groundwater), and the water flowing between the reservoirs is computed for every hydrologic response unit (HRU – the smallest computational unit for the simulations) and time step. Soil zone is simulated by three conceptual reservoirs, namely the capillary reservoir, the gravity reservoir, and the preferential-flow reservoir. These three reservoirs are not physical layers in the soil column but rather represent, and account for, soil-water content at different levels of saturation. The water contained in each of these three reservoirs is subject to different physical processes and maximum storage capacities.

ET includes five components in our PRMS models: a) evaporation of intercepted rain, b) sublimation from intercepted snow and snowpack, c) evaporation from impervious storage, d) ET from recharge zone of capillary reservoir, and e) transpiration from lower zone of capillary reservoir (Markstrom et al., 2015). Intercepted rain is assumed to evaporate at a free-water surface rate. Sublimation occurs only when there is no transpiration from plants, and sublimation loss is computed as a fraction of the potential ET (PET). The Jensen–Haise method is applied for PET calculation (Jensen et al., 1970). Note that the shrubs and trees cover types can intercept both rain and snow while the grass cover type can only intercept rain. Evaporation from the impervious portion of an HRU for each time step is based on the available water and unsatisfied PET – PET left after deducting a) and b). If the unsatisfied PET is larger (smaller) than the available water, the evaporation loss is set to the available water (unsatisfied PET). The ET term d) and e) happen if there is water in storage, and the PET demand is greater than zero after subtracting a), b), and c). These terms are handled similarly as the remaining PET demand scales by factors related to water content of the recharge zone and the capillary reservoir, respectively, and soil type.

The soil moisture variable refers to the soil water content of the capillary reservoir. Its computation is based on the summation of all moisture depletions and accretions. Depletions include ET, drainage to the groundwater reservoir, fast and slow interflow, and saturation excess surface runoff (i.e., Dunnian surface runoff). Accretions are soil infiltration and any cascading

Dunnian surface runoff and interflow from upslope HRUs. It is bounded between 0 and the maximum available capillary water-holding capacity of the soil zone.

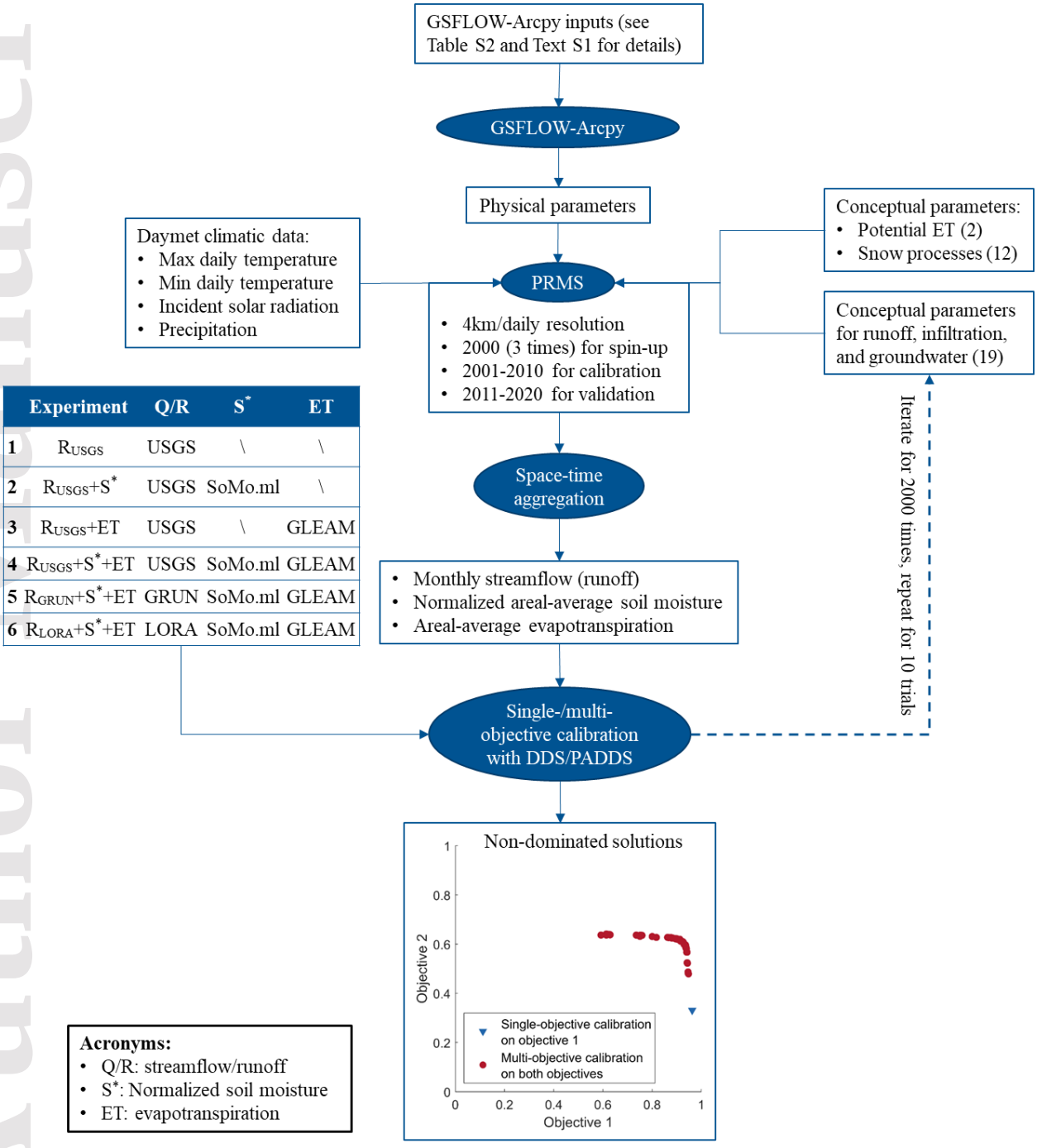
Streamflow is the sum of a) impervious Hortonian surface runoff, b) pervious Hortonian surface runoff, c) Dunnian surface runoff, d) interflow, and e) groundwater discharge (Markstrom et al., 2015). Hortonian surface runoff refers to the infiltration excess on the impervious and pervious portion of each HRU. For the impervious portion of HRU, if the sum of throughfall, snowmelt, and the antecedent impervious storage exceeds retention storage capacity for a time step, impervious Hortonian surface runoff is generated. Similarly, infiltration excess on the pervious portion of each HRU occurs when the throughfall, snowmelt, and any upslope Hortonian surface runoff available for infiltration exceed the capacity of the soil. Dunnian surface runoff and interflow are outflow from the soil zone. Excess preferential-flow reservoir inflow is the Dunnian surface runoff. The interflow consists of a fast and a slow component, which are determined by the water storage of the preferential-flow and gravity reservoir, respectively. Quadratic functions are used to model the storage-to-outflow relationship. The groundwater discharge component is assumed proportional to the groundwater storage by a coefficient.

3.2. Overview on modeling procedures

A flowchart is provided to detail the modeling procedures in **Figure 2**. Given the relatively low-relief topography of the region, a 4 km spatial discretization was chosen (i.e., every HRU is 4km-by-4km). The modeling time step is daily. The study period spans from 2000 through 2020 while the year 2000 was looped three times for model spin-up. The subsequent 10-year period (2001 to 2010) was used for model calibration and the remaining 10-year period (2011 to 2020) was used for evaluation. The Daymet maximum and minimum daily air temperature, incident solar radiation, and precipitation were resampled to 4km/daily from the original resolution and were used to force the PRMS land surface calculation. The selected modules and modeling options dictate the model parameters, which may be categorized as physical or conceptual (parameters that cannot be measured in reality). The physical parameters were derived from the datasets for model setup listed in **Table S2** in supporting information using the GSFLOW-Arcpy toolbox by Gardner et al. (2018) (**Text S1** in supporting information). The optimal values for conceptual parameters were obtained through model calibration.

There are 33 model parameters selected for calibration based on existing literature (Hunt et al., 2013; Christiansen et al., 2014). Two of the parameters are for the Jensen–Haise PET coefficients (Jensen et al., 1970). Twelve of the parameters are related to snow accumulation, melt, and sublimation. The remaining 19 parameters control surface and subsurface runoff, infiltration into the soil zone, and the rate and volume of flow from groundwater reservoirs to surface water (Markstrom et al., 2008). A table summarizing the relevant processes to these parameters and other details are provided in the supporting information (**Table S3**). These parameters were optimized in three steps following a similar approach introduced by Hay et al. (2006) and Hunt et al. (2013). The parameters calibrated in one step are then kept fixed in the subsequent steps. Note that these 33 parameters do not include any routing parameters even though the Muskingum-Manning routing scheme was adopted for the catchment models (Cunge, 1969). The reasoning for not optimizing the runoff routing parameters is provided in section 3.3.1. In calibration step one, the two Jensen–Haise parameters were optimized for a catchment; in step two, the snow processes parameters were optimized. Details on reference data and

calibration algorithms for these two steps can be found in the Supporting Information of this study. The remaining 19 parameters for runoff generation were calibrated in step three, which is the focus of this work (**Figure 2** and section 3.3.1). Coupled groundwater-surface water routing was not the purpose of our testing, therefore step four of Hunt et al. (2013) was omitted. The 20 catchment models developed with PRMS are available for downloading on ScienceBase (Mei et al., 2022).



	Experiment	Q/R	S*	ET
1	R_{USGS}	USGS	\	\
2	R_{USGS}+S*	USGS	SoMo.ml	\
3	R_{USGS}+ET	USGS	\	GLEAM
4	R_{USGS}+S*+ET	USGS	SoMo.ml	GLEAM
5	R_{GRUN}+S*+ET	GRUN	SoMo.ml	GLEAM
6	R_{LORA}+S*+ET	LORA	SoMo.ml	GLEAM

Figure 2. Flowchart of modeling procedures. Dark blue oval shapes represent the operations. Square boxes represent necessary inputs/outputs to/from the operations. The six calibration

experiments with their required observation datasets (section 3.3.1) are listed in the table. An illustration of a 2-dimensional Pareto frontier formed by a series of non-dominated solutions (red dots) is also provided (see discussion in section 3.3.2). The solution of the single-objective calibration (blue triangle) is also added.

3.3. Design of model calibration experiments

3.3.1. Model calibration schemes

There are 19 parameters for runoff generation that are obtained by model calibration. Six calibration schemes were designed with one being a single-objective and five being multi-objective problems. The six schemes involve the use of gaged streamflow and global gridded soil moisture, ET, and runoff. Experiment 1 is a single-objective calibration scheme that only considers the streamflow from USGS (R_{USGS}) to calibrate the model. Experiments 2 and 3 are bi-objective calibrations that are using gaged streamflow and either gridded soil moisture from SoMo.ml (R_{USGS}+S^*) or gridded ET from GLEAM (R_{USGS}+ET), respectively, as calibration targets. Experiment 4 is abbreviated as R_{USGS}+S^*+ET as it uses streamflow from the USGS, gridded soil moisture, and gridded ET from GLEAM. Experiment 5 and 6 are also tri-objective, but the USGS streamflow was substituted by two GRPs, namely GRUN and LORA, and are hence called R_{GRUN}+S^*+ET and R_{LORA}+S^*+ET, respectively.

The optimization algorithm applied was the Dynamically Dimensioned Search (DDS) for single-objective (Tolson, 2007) and the Pareto-Archived DDS (PADDS) for multi-objective calibration problems (Asadzadeh & Tolson, 2013), respectively. For PADDS the exact hyper-volume contribution metric was used to obtain new candidates during the calibration. For both algorithms the scalar neighborhood size perturbation parameter was fixed at its recommended value of 0.2. Both DDS and PADDS are implemented in the Optimization Software Toolkit (OSTRICH, Matott, 2016). For every experiment, a budget of 2,000 iterations was assigned to a calibration trial. In total, 10 independent calibration trials were performed for every catchment model to evaluate possible randomness of the calibration procedures. Only the best result of the 10 trials is reported here. In case of the single-objective calibration this is the best objective function value of the 10 optimal values. In case of the multi-objective calibration, the best result is the Pareto frontier derived from the pooled solutions of all 10 Pareto frontiers.

For the streamflow variable, even though the modeling time resolution is daily, the objective functions were calculated using monthly averaged values (the daily USGS streamflow and the PRMS simulated streamflow were averaged to monthly). This is to ensure a consistent comparison among different experiments that use the daily gaged streamflow and the two monthly global GRPs to inform the model. Since the monthly resolution is adopted for streamflow, the effects of runoff routing are minimized; and the monthly streamflow is essentially the monthly catchment-averaged runoff. That is the basis of using the areal-averaged GRUN and LORA to substitute the monthly gaged streamflow in experiment 5 and 6. In other words, none of the six experiments considers the effects of sub-monthly routing and associated flow, rather they focus on predicting the monthly runoff. The monthly streamflow data were used as the reference, regardless of the use of the two GRPs in experiments 5 and 6, to allow for a consistent inter-comparison among the six calibration experiments.

For soil moisture and ET, daily values were used for both the simulated and observed values. To mitigate the different spatial resolution among the SoMo.ml soil moisture, the

GLEAM ET, and the modeling resolution, the simulations and observations were areal-averaged before calculating the objective functions. For the SoMo.ml soil moisture, the volumetric soil moisture values for the three soil layers were converted to soil water depth for the entire soil column (0-50cm) before areal-averaging. In addition, although PRMS calculates volumetric soil moisture storage, it does not define an explicit surface layer depth, preventing an exact matching of soil moisture storage magnitude with the SoMo.ml derived one. Therefore, the simulated and reference soil moisture time series S_t were normalized before calculating the performance metric using:

$$S_t^* = \frac{S_t - \mu_S}{\sigma_S} \quad . \tag{1}$$

where μ_S and σ_S represent the mean and standard deviation of soil moisture. Eq.(1) means that the magnitude and variability of soil moisture from SoMo.ml were not used to inform the parameter optimization process while only the signal's timing was considered.

3.3.2. Calibration metrics

In all six calibration schemes, the Kling-Gupta Efficiency (KGE; Gupta et al., 2009) is used to evaluate the performance of the hydrological variable simulations:

$$KGE_x = 1 - \sqrt{\left(\frac{\mu_x^m}{\mu_x^o} - 1\right)^2 + \left(\frac{\sigma_x^m}{\sigma_x^o} - 1\right)^2 + (r_x - 1)^2} \quad , \tag{2}$$

where μ_x and σ_x represent the mean and the standard deviation of the time-dependent variable x, which can be streamflow (Q), normalized soil moisture (S^*), or ET. The superscript m and o indicate modeled and observed time series, respectively. The Pearson correlation coefficient r_x is derived between the modeled and observed time series of the respective variable x. The KGE_x is bound by (-∞, 1] with 1 being the ideal value. For the normalized soil moisture, the mean is 0 and the variance is 1 for both the observed as well as the modeled time series. Hence, the bias terms for the mean and the standard deviation, the first and the second term in Eq.(2), are essentially 0. Therefore, Eq.(2) collapses to

$$KGE_{S^*} = r_{S^*} \quad . \tag{3}$$

Eq.(3) indicates that only the time information of the reference soil moisture is used to inform the model simulation.

Instead of merging the multiple objective functions into one, we maintain several separated objective functions and perform multi-objective calibration. Unlike single-objective calibration that obtains a single optimal solution, multi-objective calibration identifies a set of non-dominated solutions (NDSs) that forms a Pareto frontier. A solution is non-dominated if none of its objective functions can be improved without degrading some of its other objectives. **Figure 2** demonstrates a Pareto frontier for a 2-dimensional (two-objective) problem (red circles), while in our study the dimension can be up to three, i.e., one for streamflow, soil moisture, and ET. The front might reveal that there are solutions where introducing an additional objective can lead to a reduced performance of the original objective (compared to blue marker), while the performance of the additional objective increases.

3.4. Model performance evaluation

3.4.1. Performance metrics and standards

The KGE_x metric used in model calibration and a combined efficiency metric that determines the overall model performance across all variables are used to quantify the model performance during the 2011-2020 evaluation period. The USGS streamflow, SoMo.ml soil moisture, and GLEAM ET are used as the reference to be consistent with the calibration period. The combined efficiency metric CE_{eq} defined as the arithmetic mean of the performances for each of the three variables is introduced:

$$CE_{eq} = \frac{KGE_Q + KGE_{S^*} + KGE_{ET}}{3} \quad . \tag{4}$$

The subscript *eq* indicates that the weighting factors for the three objective functions are the same (here: $\frac{1}{3}$). The range of CE_{eq} is (-∞,1] with 1 as the ideal value, the same as the KGE's.

To help interpreting the error metrics, we defined three model performance standards for KGE_x that are gradually decreasing based on the goodness of the simulations following Mai et al. (2022). Consider the three terms within the square root sign of Eq.(2); the first level is characterized by within 10% under-/over-estimation of the observation in terms of the mean ($|\frac{\mu_x^m}{\mu_x^o} - 1| \leq 10\%$) and the SD ($|\frac{\sigma_x^m}{\sigma_x^o} - 1| \leq 10\%$) and the correlation coefficient between model output and observation is above 0.9 ($r_x \geq 0.9$). These criteria together result in a KGE no less than 0.83. By gradually releasing these criteria, we also define another two performance levels with lower upper boundaries of KGE values at 0.65 and 0.48, respectively. These performance levels are applicable for the CE_{eq} metric in Eq.(4) and a more general form that will be introduced in Eq.(6), section 3.4.3.

3.4.2. Pairing experiments for relative performance

Objectives A and B of this study are addressed by comparing relative performance among the six calibration experiments. To quantify the improvement/deterioration in model performance between two calibration experiments *i* and *j*, the median performance difference of all pairs of NDSs from experiment *i* and *j* for a selected variable ($\tilde{\Delta}_X$) is derived:

$$\tilde{\Delta}_X = med\langle X_i - X_j \rangle \quad with \quad 1 \leq j < i \leq 6 \quad . \tag{5}$$

The term *X* is one of the error metrics defined in Eq. (2) to (4); subscripts *i* and *j* are the experiment numbers ranging from 1 to 6. Note that the difference is always the experiment with the higher number (*i*), minus the experiment with the lower number (*j*). If the Pareto frontier of experiment *i* contains *M* NDSs and experiment *j* has *N* NDSs, $\tilde{\Delta}_X$ is defined as the median of the $M \times N$ differences that can be calculated. A positive (negative) value of $\tilde{\Delta}_X$ means that, for a catchment model, experiment *i* is better (worse) than experiment *j* in terms of the median of all possible pairwise differences from NDSs of the two experiments.

3.4.3. Visualizing model performance by ternary diagram

Objective C of this study is to investigate the model performance dynamics given different importance of the three variables. To this end, a method that effectively synthesizes the weighting factor spaces with all the NDSs of a catchment model in a ternary diagram (Howarth, 1996) is proposed. The NDSs of a catchment model are obtained by merging those from the six calibration experiments, i.e., identifying the set of NDSs when the six sets of NDSs are merged. For each of the overall NDSs, we introduce a more general combined efficiency metric after model calibration, *CE*, with weighting factors compared to CE_{eq} in Eq.(4):

$$CE = w_Q KGE_Q + w_{S^*} KGE_{S^*} + w_{ET} KGE_{ET} \quad , \tag{6}$$

where w_Q, w_{S*}, and w_{ET} are the weighting factors for streamflow, normalized soil moisture, and ET, respectively. For each of the NDS, *CE* is a function of the three weighting factors bounded by (-∞,1] as this is the range of the three individual metrics. Given that w_Q, w_{S*}, and w_{ET} are bounded by [0,1] and sum to 1, the values of *CE* under the possible weighting factor space can be represented on a ternary diagram with the three weighting factors being the three axes. We called this a "CE surface."

For a catchment with *N* NDSs, there are also *N* CE surfaces. Visualizing all of these CE surfaces would also require *N* ternary diagrams, which is not efficient if one would like to investigate the change of model performance indicated by CE over the possible weighting factor spaces. Hence, the median over the stack of the CE surfaces is defined as an aggregation:

$$\widetilde{\boldsymbol{CE}} = med\langle \boldsymbol{CE} \rangle \quad , \tag{7}$$

where $\boldsymbol{CE}$ is the stack of the CE surfaces each defined through Eq.(6). $\widetilde{\boldsymbol{CE}}$ is the median performance of a model across differently weighted objectives; its range is (-∞,1], the same as *CE*. Note that the stack of CE surfaces now collapses to a single surface represented by $\widetilde{\boldsymbol{CE}}$, and this surface is visualized through a ternary diagram. There is one such diagram for each catchment.

4. Results

4.1. Overall performance of experiments

To analyze the nominal performance of each of the six experiments, the four performance metrics for all the NDSs across all 20 catchments are summarized as boxplots in **Figure 3**. In general, using the datasets R_{USGS}, R_{USGS}+S^*, and R_{USGS}+ET for model calibration leads to the best model performances regarding streamflow (**Figure 3**a), soil moisture (**Figure 3**b), and ET (**Figure 3**c), respectively. This is consistent with most of the studies shown in **Table 1**: introducing a new variable as an additional objective reduces the streamflow simulation performance but improves the performance of the additional variable's simulation in return. In terms of the overall performance, the single-objective and the two bi-objective calibration schemes result in generally better performance than the tri-objective R_{USGS}+S^*+ET as revealed by the CE_{eq} distribution (**Figure 3**d). A closer look at the results of the first three calibration schemes in **Figure 3**a reveals that the decrease in KGE_Q of adding S^* as an additional model calibration constraint is not so severe as adding ET. A cross-benefit is identified through the result shown in **Figure 3**c where introducing soil moisture in the calibration also improves ET

(compare whiskers of boxes for R_{USGS} and R_{USGS}+S^*); this is not the case when ET information is used in addition to streamflow (compare whiskers of R_{USGS} and R_{USGS}+ET in **Figure 3**b).

Among the three experiments that calibrated against three variables, the gaged streamflow-based calibration scheme outperforms the two gridded runoff-based ones in terms of KGE_Q and CE_{eq}. This is because the USGS streamflow is used as reference to calculate the performance metrics for all experiments. More fair evaluations could be to compare the soil moisture and ET simulations among the three experiments. In fact, **Figure 3**b and **3**c reveal that the R_{GRUN}+S^*+ET and R_{LORA}+S^*+ET soil moisture and ET simulations are almost identical to the R_{USGS}+S^*+ET ones. A product-wise comparison shows that R_{GRUN}+S^*+ET is better than R_{LORA}+S^*+ET for streamflow simulation (**Figure 3**a) while being fairly similar regarding soil moisture (**Figure 3**b) and ET (**Figure 3**c). This leads to an overall better model performance when using GRUN instead of LORA in model calibration (**Figure 3**d).

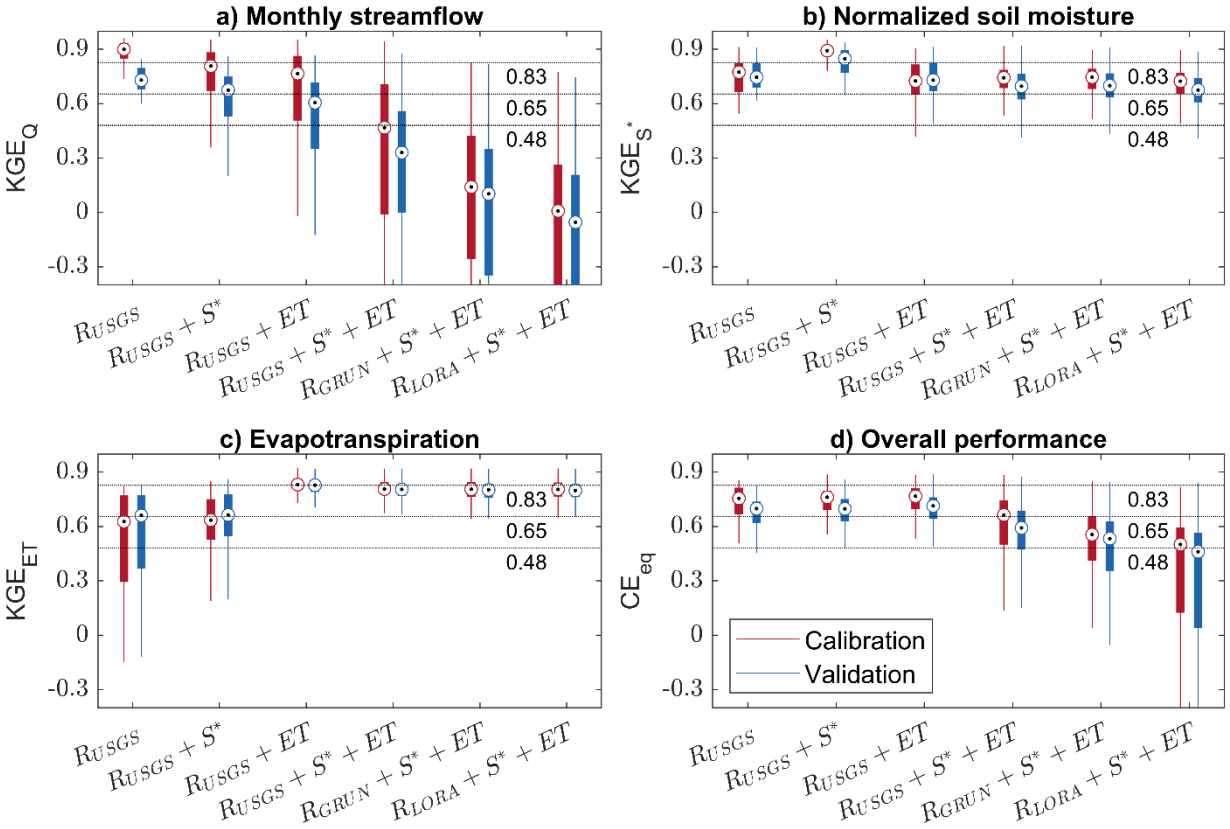


Figure 3. Performance metrics for a) monthly streamflow, b) normalized soil moisture, c) ET, and d) overall performance for the six experiments defined in section 3.3.1. The boxplot shows the results for all 20 catchments and all NDSs. The horizontal lines from top to bottom correspond to the three performance levels (section 3.4.1**)**.

4.2. Relative performance between experiments

The relative performances among experiments 1 to 4 across all the 20 catchments are summarized in **Figure 4**. The figure shows the number of catchments where the median performance of the model improved with adding/substituting a variable compared to another

experiment. A large number of catchments (blue colors) indicates a reliable improvement of model performance with a specific variable being added/substituted. The metric that is used to quantify the performance of the model is added as a label in each grid.

Focusing on the grids that compare R_{USGS}+S^* with R_{USGS} for KGE_{ET}, 12 catchments show improvements in model performance regarding ET after adding soil moisture as an additional objective for both the calibration and evaluation period. This confirms the cross-benefit for 60% of the catchments of using soil moisture in model calibration to improve the performance of ET simulations (**Figure 3**c). The cross-benefit of ET data to improve the quality of soil moisture simulations is less clear. Focusing on the grids comparing R_{USGS}+ET to R_{USGS}, there are 8 (11) catchments that show an increase in KGE_{S*} with the addition of ET in calibration for the calibration (evaluation) period. In addition, by comparing the R_{USGS}+S^* and the R_{USGS}+ET schemes, it is revealed that 8 catchments that show higher KGE_Q values for R_{USGS}+ET compared to R_{USGS}+S^* in the calibration as well as the evaluation period. This indicates a lower degree of degradation in performance regarding streamflow simulations with the addition of soil moisture data rather than ET data. Lastly, the results show that no more than 4 catchments yield a higher CE_{eq} for the R_{USGS}+S^*+ET compared to all the other single- and bi-objective calibration schemes in both calibration and evaluation.

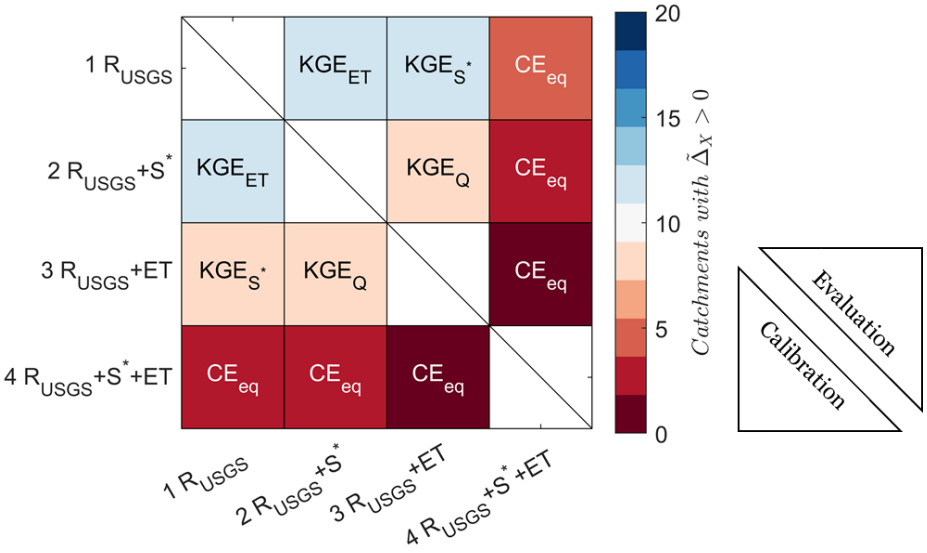


Figure 4. Number of catchments that show improvements in model performance when adding additional variables (soil moisture, ET, or both), i.e., the median of the error metric differences $\tilde{\Delta}_X$ is larger than 0. The metric considered to compare each pair of experiments is added as a label in each grid. The upper (lower) triangle is estimating model performance of the evaluation (calibration) period.

The performance among experiments 4 through 6 are analyzed to estimate the impact of replacing the gaged streamflow (experiment 4) with GRPs (i.e., GRUN in experiment 5 and LORA in experiment 6) during model calibration. **Figure 5** displays the results across the 20 study catchments. **Figure 5**a shows that in no more than 7 catchments the model performance regarding streamflow (KGE_Q) is improved when the model is trained with a GRP (R_{GRUN}+S^*+ET and R_{LORA}+S^*+ET) instead of USGS streamflow (R_{USGS}+S^*+ET); the remaining catchments show a decrease in performance. **Figure 5**d shows a similar pattern regarding overall model performance (CE_{eq}), i.e., no more than 8 catchments show an improved overall performance

when trained with GRPs while the remaining catchments show a decrease in performance. This agrees with the observations from **Figure 3**d that R_{USGS}+S^*+ET outperforms R_{GRUN}+S^*+ET and R_{LORA}+S^*+ET in terms of KGE_Q and CE_{eq}.

Figure 3 already had revealed that the three tri-objective calibration schemes (experiments 4 to 6) yield almost identical performance in terms of KGE_{S*} and KGE_{ET}. Yet, using the number of catchments with positive $\tilde{\Delta}_X$, i.e., the number of catchments being improved, some nuances in performance between the experiments are identified. For instance, the improvements in KGE_{S*} values of R_{GRUN}+S^*+ET are higher than those of R_{USGS}+S^*+ET for 13 (16) catchments for the calibration (evaluation) period (**Figure 5**b), while the KGE_{ET} values of R_{LORA}+S^*+ET are higher than the R_{USGS}+S^*+ET ones for 11 (12) catchments for the calibration (evaluation) period (**Figure 5**c).

To understand the relative performance between R_{GRUN}+S^*+ET and R_{LORA}+S^*+ET, we investigate all the R_{GRUN}+S^*+ET vs. R_{LORA}+S^*+ET grids in **Figure 5**. The results show that using the LORA runoff product instead of GRUN leads to improvements regarding ET in 15 catchments (**Figure 5**c) while the use of the GRUN data leads to better results regarding streamflow, soil moisture, and overall performance (**Figure 5**a, **5**b, and **5**d, respectively). Specifically, only 1, 3, and 1 catchments show better performance with respect to streamflow, soil moisture, and overall performance, respectively, when the LORA runoff is used instead of GRUN. These numbers of catchments are the same for both the calibration and the evaluation period.

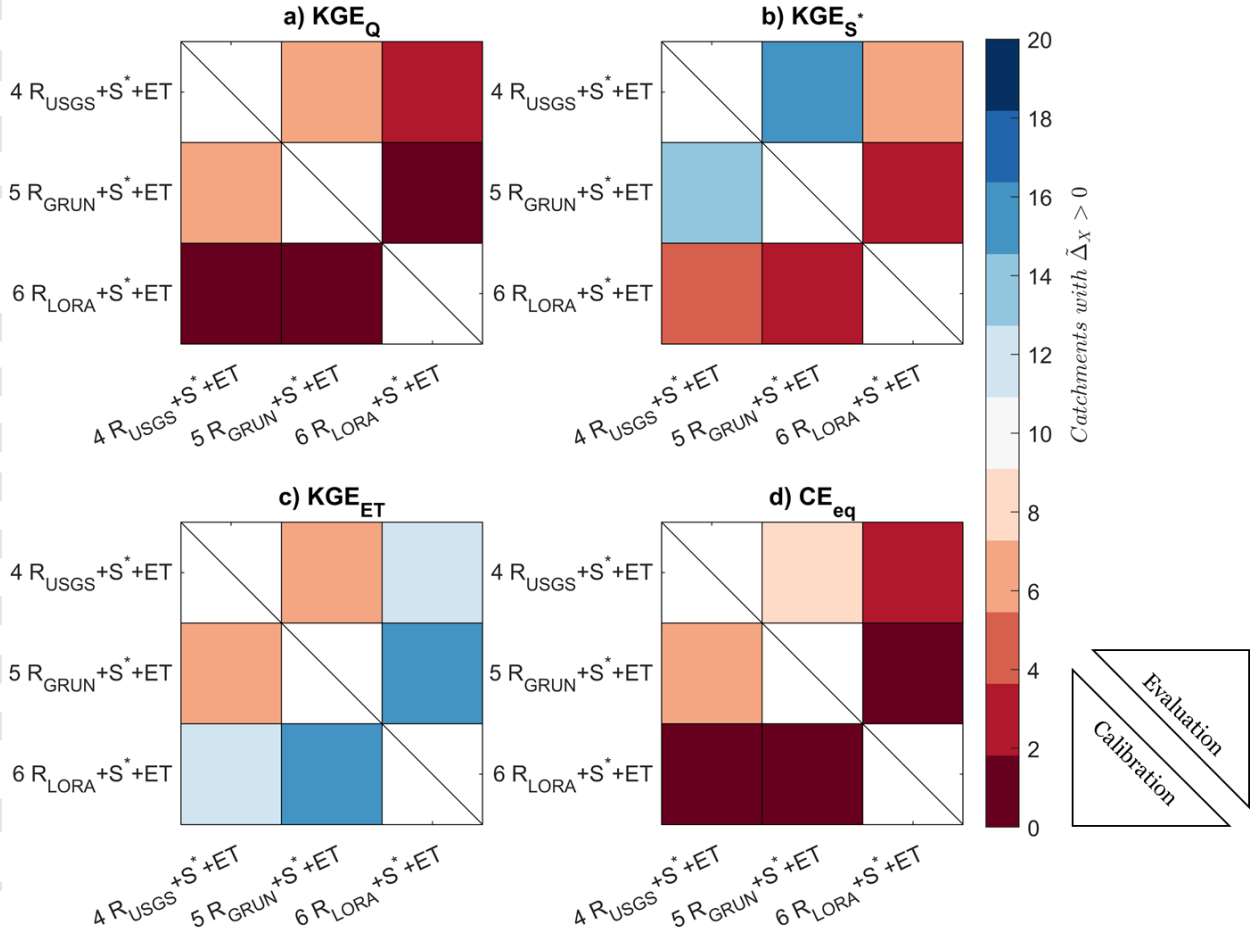


Figure 5. Number of catchments that yield an improvement in model performance when the gaged streamflow data (R_{USGS}) are replaced by gridded runoff products (R_{GRUN} and R_{LORA}), i.e., the median of the error metric differences $\tilde{\Delta}_X$ is larger than 0. The different panels show the result considering a) KGE_Q, b) KGE_{S*}, c) KGE_{ET}, and d) CE_{eq} as the metric to determine the model performance. The upper (lower) triangle is estimating model performance of the evaluation (calibration) period.

4.3. Relative performance of global runoff products on catchment scale

The impact of replacing gaged streamflow observations with GRPs on model performance is further analyzed with respect to catchment area. **Figure 6** shows the median improvement regarding streamflow performance $\tilde{\Delta}_{KGE_Q}$ when comparing the baseline experiment using USGS streamflow for model calibration vs. the two experiments using either GRUN (**Figure 6**a) or LORA (**Figure 6**b). The figure shows the individual improvements of 19 catchments excluding catchment g, which is identified as an outlier (see section 4.4 and **Figure 7**g). The $\tilde{\Delta}_{KGE_Q}$ values are mostly negative, indicating that models calibrated to gaged streamflow yield better simulations compared to models calibrated to either of the two GRPs. This is consistent with the results shown in **Figure 5**a. Both panels of **Figure 6** suggest a positive correlation between the performance improvements $\tilde{\Delta}_{KGE_Q}$ and the catchment area. This indicates that the streamflow simulation discrepancies for models that are calibrated using GRPs compared to models that are trained using gaged streamflow data are mitigating from small to large scale catchments. The correlation is larger for experiments using the GRUN dataset (**Figure 6**a) which means that the scale dependency is more obvious for the GRUN dataset than when the LORA dataset is used.

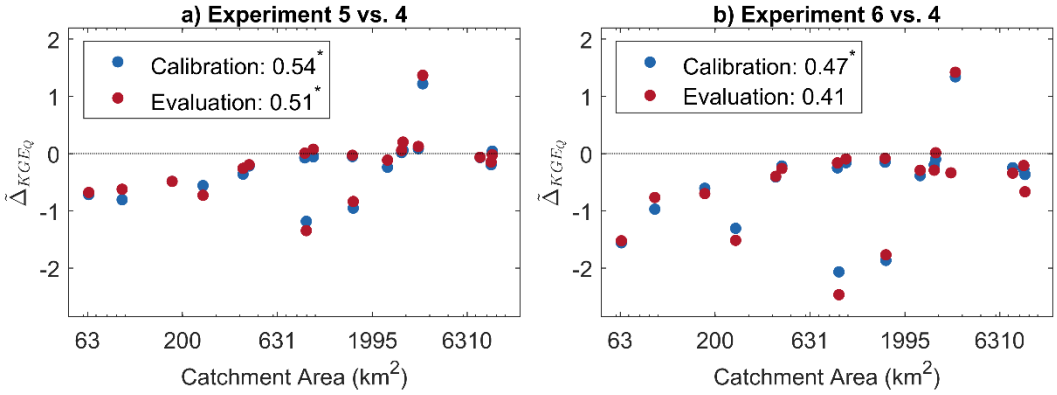


Figure 6. Model performance changes regarding streamflow ($\tilde{\Delta}_{KGE_Q}$) when replacing USGS streamflow data (experiment 4) with gridded runoff data from a) GRUN (experiment 5) or from b) LORA (experiment 6) dependent on the catchment drainage area. The performance changes are estimated for the calibration (blue markers) and evaluation (red markers) period. The correlation coefficient between the streamflow performance changes and the catchment area are added to the legend; an asterisk * is added in case the statistical significance is at least at the level of 0.05.

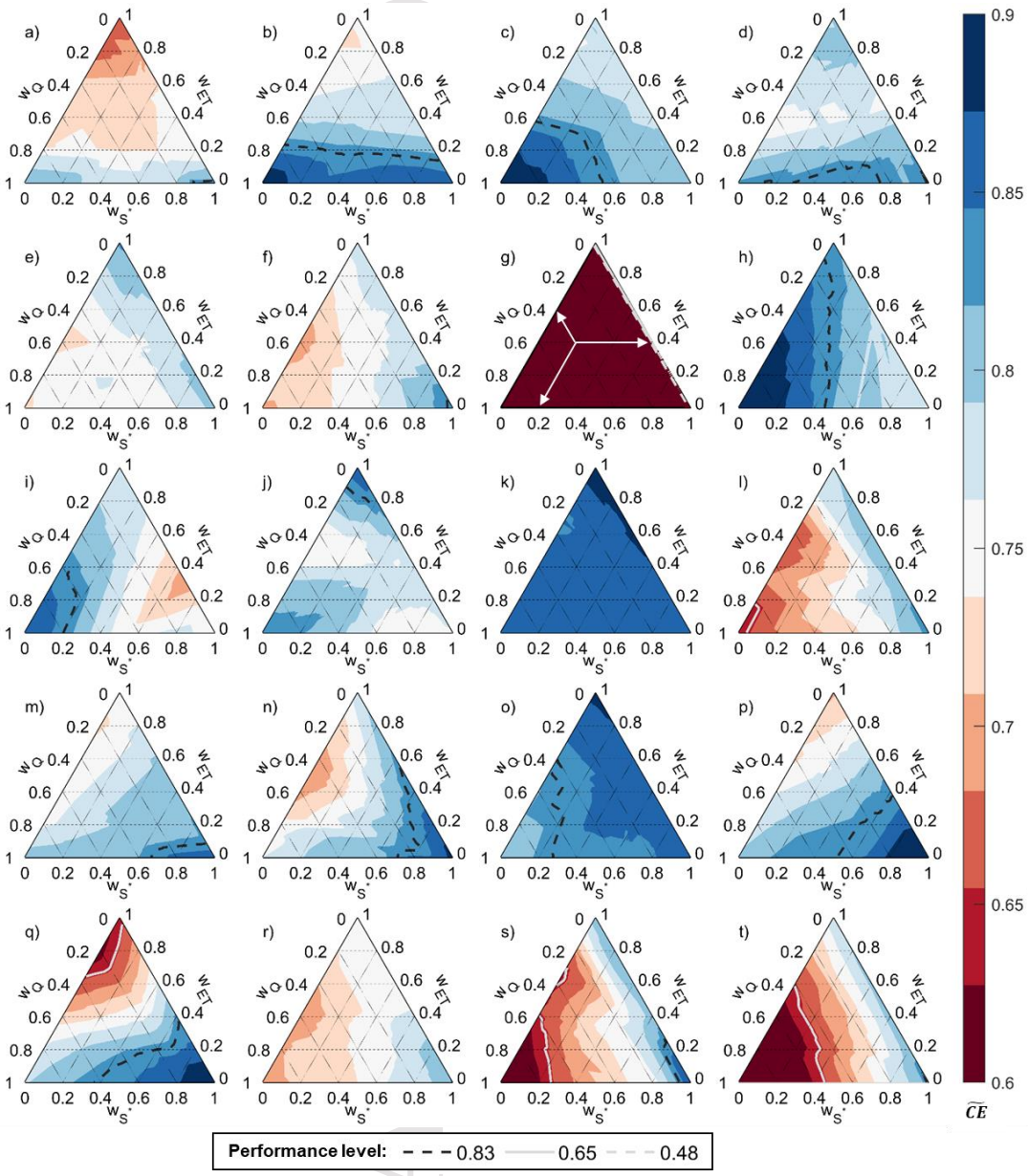
4.4. Dynamics of overall modeling performance on variable importance

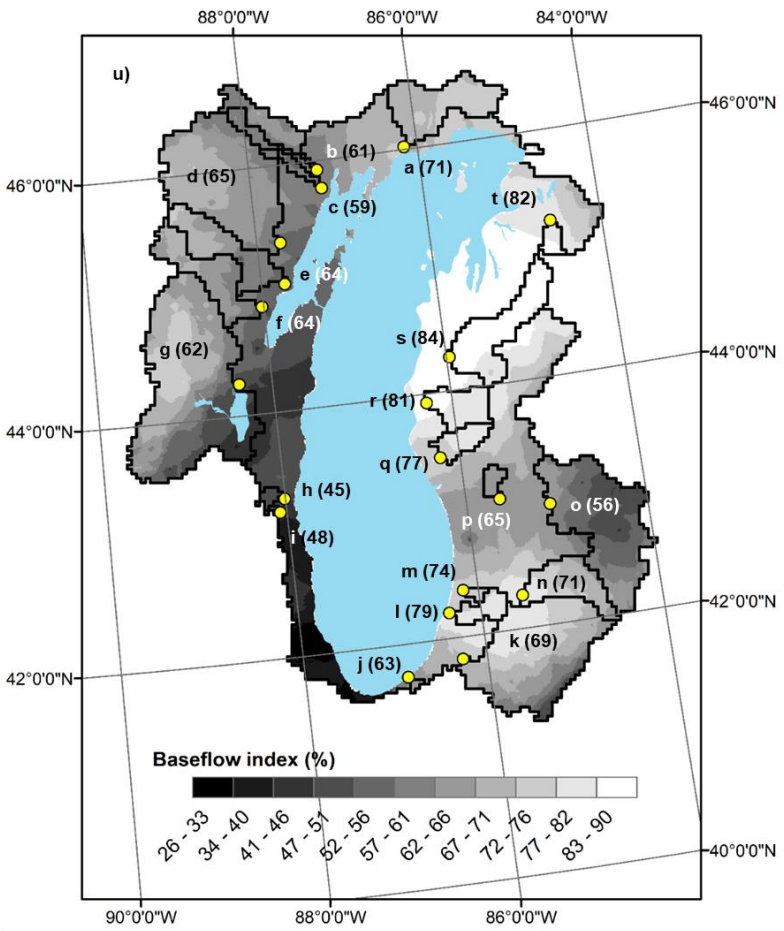
To demonstrate the model performance as a function of weighting factors assigned to the objective functions after model calibration, the $\widetilde{\boldsymbol{CE}}$ metric for the calibration period is shown as ternary diagrams for the 20 catchments in **Figure 7** (the 20 $\widetilde{\boldsymbol{CE}}$ surface for the evaluation period are provided in **Figure S3**). A baseflow index map (Wolock, 2003) illustrating the spatial distribution of baseflow as a percentage of total streamflow is added for reference to discuss model performance varied on baseflow contributions. Note that the hydrological model for catchment g (**Figure 7**g) is problematic due to the highly regulated streamflow regime by reservoir operations and the fact that no reservoir management rules were available and used to build the hydrological model. This leads to very low KGE_Q values for catchment g that impact any overall performance estimate $\widetilde{\boldsymbol{CE}}$ even if the streamflow performance metric is weighted very low. A performance level of $\widetilde{\boldsymbol{CE}} \geq 0.48$ can only be achieved if the weight of the streamflow performance w_Q is as low as 1.8% leading to an entirely dark red colored ternary plot for catchment g. All other catchments show $\widetilde{\boldsymbol{CE}}$ gradients from 0.49 to 0.92 depending on the weights chosen for each component.

The details of each $\widetilde{\boldsymbol{CE}}$ surface are investigated by focusing on the trend and the magnitude. It can be seen from **Figure 7** that some catchments show unified $\widetilde{\boldsymbol{CE}}$ patterns. For example, catchments k and o reveal mild variation of their $\widetilde{\boldsymbol{CE}}$ surfaces, which are attributed to their relatively good performance for all three variables ($\widetilde{\boldsymbol{CE}} > 0.79$). Catchments l, s, and t have among the weakest streamflow simulations; their $\widetilde{\boldsymbol{CE}}$ contour lines are approximately parallel to the w_{ET} axis, showing a decreasing trend with increasing streamflow weights and mild gradient at the direction that the streamflow weight is fixed. These observations pinpoint that soil moisture and ET are simulated equally well and are better than their respective streamflow simulations for the three catchments. It is worth noting that the catchments l, s, and t are also associated with the highest mean baseflow indexes across the watershed (79%, 84%, and 82%, respectively), indicating high contributions of baseflow to their streamflow. A similar situation can be observed for catchment r with the third highest baseflow index (81%). This finding of weak streamflow simulation performances for catchments that are in regions with high baseflow indexes is confirming results found by Fry et al. (2014). Other similar trends can be observed from the catchment groups (a, b) and (p, q). For the first group, their $\widetilde{\boldsymbol{CE}}$ contour lines are approximately parallel to the w_{S^*} axis, decreasing towards the vertex of the triangles. This indicates weakest ET simulation among S^* and Q, whose performances are similar. The second group is similar to the first group as their weakest ET simulations. But the $\widetilde{\boldsymbol{CE}}$ contour lines of these two catchments are not parallel to the w_{S^*} axis, twisted counterclockwise for some degree, as S^* are simulated obviously better than Q.

For magnitude of $\widetilde{\boldsymbol{CE}}$, if one would use a dominant performance level that covers more than 50% of the ternary diagram's surface area, the majority of 17 catchments reach the performance level of $\widetilde{\boldsymbol{CE}} \geq 0.65$. The model for catchment k and o are the only two that are dominated by the highest level of $\widetilde{\boldsymbol{CE}} \geq 0.83$. The results are summarized under "Overall" in the table added to **Figure 7**. We also analyze the three corners of each triangle in order to identify which variable is simulated best in each catchment. We define a corner to be the part of the triangle where the weights are larger than 80% for one variable. This assessment is meaningful for when one believes a particular variable is subjected to notably less uncertainty (take the ground-based streamflow measurements vs. other global gridded products as an example). The

dominant performance level of each corner (performance level covers more than 50% of the area) are counted for the 20 catchments and results are summarized in the table added to **Figure 7**. The soil moisture simulations exhibit overall the best performance with seven catchments being at the level of $\widetilde{\boldsymbol{CE}} \geq 0.83$. While it is hard to clearly state whether streamflow or ET is simulated with a better overall performance across catchments based on the distribution of the cases.





	Overall	Streamflow	Soil moisture	Evapotranspiration
[0.83, 1]	2 (10%)	5 (25%)	7 (35%)	3 (15%)
[0.65, 0.83)	17 (85%)	12 (60%)	12 (60%)	15 (75%)
[0.48, 0.65)	0 (0%)	2 (10%)	0 (0%)	1 (5%)
(-∞, 0.48)	1 (5%)	1 (5%)	1 (5%)	1 (5%)

Figure 7. Median of the combined performance metric $\widetilde{\boldsymbol{CE}}$ using different weights for each of the three objectives across the 20 catchments (panel a to t) for the calibration period. A demonstration on how to read the weighting factors for a point on the $\widetilde{\boldsymbol{CE}}$ surface is provided in panel g. Lines of different gray scales are added to the ternary plots to visualize the three performance levels (section 3.4.1). Panel u) shows a map of the baseflow indexes for the Lake Michigan watershed as well as the location and extent of the twenty catchments. The labels of the catchments (a-t) correspond to the panels (a to t). The catchment-averaged baseflow indices are reported within the parenthesis. The table summarizes the distributions of model performance.

5. Discussion

5.1. On relative performance among adding variables for model calibration

Our analysis reveals that using both normalized soil moisture and ET as additional variables to augment streamflow in model calibration (the tri-objective calibration scheme) results in the lowest overall modeling performance (*CE* in Eq.(4)) compared to the single- and the two bi-objective schemes. There are two possible reasons for seeing a decreasing overall performance: First, the model structure may not fully resolve the physical processes and hence it may be impossible to satisfy all three constraints at the same time. Second, the data products used for calibration may not accurately capture the magnitude and seasonality of the natural processes. These explain why streamflow, soil moisture, and ET cannot be matched at the same time.

Another important result of this work is that there is a cross-benefit of including soil moisture in model calibration (in addition to streamflow) to the ET simulation for most of the catchments. This may be explained as the improved representativeness of water storage fluctuations of the recharge zone and the lower zone of the capillary reservoir, which in turn benefits the simulations of the associated ET fluxes. Note that soil moisture is the water content of the capillary reservoir (section 3.1). Therefore, the more accurate soil moisture simulation by including the variable in calibration as adopted by the $R_{USGS+S*}$ scheme yields better recharge zone ET and lower soil zone transpiration compared to the R_{USGS} scheme. On the contrary, fewer catchments show that the addition of ET in calibration benefits soil moisture. This is probably because the five ET components (see section 3.1 for details) were considered together in model calibration, which does not solely inform the parameters of the two soil moisture related components. Another benefit of adding soil moisture compared to adding ET is that the former one reduces streamflow performance less. This could be because only the timing of soil moisture is used in addition to streamflow to inform the model parameters, while all the information of the ET time series is used for the other two schemes.

5.2. Global runoff vs. gaged streamflow in model calibration

This study renders the first ever comparison of two global runoff products, namely GRUN and LORA, in hydrological model calibration. This is the first ever study that uses GRUN in process-based hydrological model calibration to the best of our knowledge. Our results demonstrate that the gaged streamflow-based calibration outperforms the global product-based ones in terms of streamflow simulation. This is not unexpected given that the USGS streamflow is adopted as one of the reference datasets in training the data-driven models to produce GRUN and LORA (Ghiggi et al., 2019; Hobeichi et al., 2019). Note that GRUN is trained to the Global Streamflow Indices and Metadata Archive (GSIM), which contains monthly streamflow records from 9404 USGS streamflow stations (Do et al., 2018). So, the discrepancy between GRUN/LORA and the gaged streamflow propagates to the streamflow simulation through model calibration. Our results also indicate that as the catchment scale increases, the streamflow simulation discrepancy is mitigated by the global product-based models and gaged streamflow-based models. This is probably because of the discrepancy in scale between the two gridded runoff products (~50km) and the small-scale catchments analyzed herein.

Another observation regarding the relative performance between GRUN and LORA is that the GRUN-calibrated models had superior performance than the LORA counterparts in terms of streamflow simulation. Given the fundamentally different algorithms of GRUN and LORA, the relative differences between GRUN and LORA could be attributed to a wide range of factors. Yet, a clear reason that could partly explain this observation is the different number of streamflow gages that were used to produce GRUN and LORA. By visually inspecting the gage density over the Great Lake region for GRUN (Figure 2 in Ghiggi et al. (2019)) and that for LORA (Figure 1 in Hobeichi et al. (2019)), one could discover that the former one has more gages, which may lead to lower discrepancy to the reference network.

5.3. Diagnosing the dynamics of model performance on variable importance

This study uses the ternary diagram to visualize the performance of three hydrological variables (streamflow, soil moisture, and ET). It shows how the overall modeling performance (the $\widetilde{\boldsymbol{CE}}$ metric in Eq.(7)) change on the variable importance space. The visualization method is also flexible to be applied for other hydrological models and other flux and state variables. This is meaningful for the model's end-user with different emphasis on the hydrological simulations (e.g., flow simulations, planning of agriculture activities). With the diagnostic information provided by the $\widetilde{\boldsymbol{CE}}$ surface, the modelers can refine the processes that they are interested in and are less satisfied with for their modeling practices. Other error metrics may also be used to construct the $\widetilde{\boldsymbol{CE}}$ surface with a different formula for averaging. For instance, the geometry distance to the utopian point (the point that all variables reach the ideal performance level) is also a popular metric for representing the overall modeling performance; see for example Herman et al. (2018).

5.4. Strategies to improve streamflow simulation for baseflow-dominated catchments

Results show that the streamflow simulation is relatively weak for baseflow-dominated catchments (**Figure 7**). A potential cause could be the simple representation of groundwater in PRMS that does not fully resolve the subsurface processes; groundwater discharge is assumed proportional to groundwater storage by a coefficient (section 3.1). Therefore, one way to improve streamflow simulation by PRMS could be to replace this simple reservoir conceptual model by MODFLOW which is available under the GSFLOW platform, or an equivalent physically based, spatially distributed, groundwater model. Inclusion of the groundwater system allows integration of groundwater observations (e.g., groundwater head, water table depth), which can then inform the model's ability to simulate Dunnian surface runoff – an important process for forecasting peak flows. That is, similar to the ability of soil moisture and ET data to constrain related watershed hydrological components, the addition of groundwater processes will improve the groundwater recharge component, something only indirectly informed using streamflow or other hydrological variables that only reflect the land surface processes (Huntington & Niswonger, 2012; Xu et al., 2021). Indeed, the calibration of the GSFLOW model of Hunt et al. (2013) used a wide variety of observation types, including snow depth, lake evaporation, actual ET, streamflow, lake stage, groundwater levels, groundwater inflows to lakes, and depth of lake plumes to constrain the parameters used to simulate the watershed hydrological components. In this way the results of our work are consistent with others who have noted the value of a wide variety of data types for watershed flow calibration (e.g., Hunt et al., 2006).

6. Conclusions

In this study, we conducted six model calibration experiments using different observation data sets, including gaged streamflow and global gridded products of soil moisture, evapotranspiration, and runoff over 20 catchments located in the Lake Michigan watershed. The soil moisture, evapotranspiration, and streamflow simulations produced from the six experiments were compared. A novel model performance visualization method was presented using a combined efficiency metric and the ternary diagram. Our results suggest that, among the six experiments, the single- and bi-objective calibration schemes yielded the best overall modeling performance; the addition of soil moisture improves prediction of evapotranspiration for most of the catchments due to correlation between the two variables.

Regarding the potential of using global gridded runoff products in model calibration, we found that models informed by gaged streamflow outperform the gridded runoff product counterparts. This is because the simulated streamflow inherits the discrepancy between the global runoff products and the gaged streamflow. However, as the spatial aggregation scale for the global runoff products increases for larger catchments, the difference between the two types of models diminishes. For soil moisture (evapotranspiration), most of the GRUN- (LORA-) based models show better performance than the USGS streamflow-based ones. Between the two products, we found that GRUN-informed models provide better streamflow and soil moisture simulations than the LORA counterparts, while the LORA-based models are generally better than the GRUN ones for evapotranspiration.

According to the ternary diagram, some typical trends are identified. The simulations of the normalized soil moisture show better performance than the streamflow and the evapotranspiration ones for most of the catchments. Relatively low streamflow simulation performance is found for catchments with high baseflow contribution.

Acknowledgements

This work was conducted as a part of the “Improving representation of groundwater in foundational Great Lakes hydrologic and hydrodynamic models and data sets” Working Group supported by the John Wesley Powell Center for Analysis and Synthesis, U.S. Geological Survey. Support for Niswonger, Regan, and Hunt was provided through the U.S. Geological Survey Water Availability and Use Program. The authors are thankful to Drs. Melissa D. Masbruch, Jesse E. Dickinson, and Andy Bock for their technical supports and useful suggestions. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data Availability Statement

The HydroSHEDS DEM data are available from https://www.hydrosheds.org/downloads. The NALCMS land cover dataset is downloaded from http://www.cec.org/north-american-environmental-atlas/land-cover-2010-modis-250m/. The GMIS v1 impervious surface dataset is freely available from https://sedac.ciesin.columbia.edu/data/set/ulandsat-gmis-v1/data-download. The MODIS MOD44B vegetation characteristics product is downloaded from the NASA Earthdata Search website https://search.earthdata.nasa.gov/search. The SoilGrids v2 soil characteristics datasets are downloaded from https://soilgrids.org/. The Daymet v4 dataset is available from https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1840. The GHCN-Daily data are downloaded from https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily. The USGS streamflow data are downloaded from https://waterdata.usgs.gov/nwis/dv/?referred_module=sw. The SoMo.ml v1 soil moisture datasets are retrieved from https://www.bgc-jena.mpg.de/geodb/projects/Data.php. The GLEAM v3.5b dataset is retrieved from https://www.gleam.eu/#downloads. The GRUN v1 runoff product is available from https://figshare.com/articles/dataset/GRUN_Global_Runoff_Reconstruction/9228176. The LORA v1 runoff product is downloaded from https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f9617_9854_8096_5291. The GSFLOW model v2.2.0 is freely available from https://water.usgs.gov/water-resources/software/gsflow/. The GSFLOW-Arcpy toolbox is downloaded from https://github.com/gsflow/gsflow-arcpy. The model calibration software Ostrich is available from https://www.eng.buffalo.edu/~lsmatott/Ostrich/OstrichMain.html. The model archive for

the 20 study catchments will be publicly available at the end of this project at https://doi.org/10.5066/P9DOVISZ.

References

Asadzadeh, M. & Tolson, B., 2013. Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization. *Eng. Optimiz.*, 45(12), pp. 1489-1509.

Bai, P., X, L. & Liu, C., 2018. Improving hydrological simulations by incorporating GRACE data for model calibration. *J. Hydrol.,* Volume 557, pp. 291-304.

Beck, H. E. et al., 2016. Global-scale regionalization of hydrologic model parameters. *Water Resour. Res.*, 52(5), pp. 3599-3622.

Beven, K., 2006. A manifesto for the equifinality thesis. *J. Hydrol.*, 320(1), pp. 18-36.

Brown de Colstoun, E. C. et al., 2017. *Documentation for the Global Man-made Impervious Surface (GMIS) Dataset From Landsat,* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Champagne, O. et al., 2020. Future shift in winter streamflow modulated by the internal variability of. *J. Hydrol.*, 24(6), pp. 3077-3096.

Christiansen, D. E., Walker, J. F. & Hunt, R. J., 2014. *Basin-scale simulation of current and potential climate changed hydrologic conditions in the Lake Michigan Basin, United States,* U.S. Geological Survey Scientific Investigations Report 2014-5175.

Cunge J. A., 1969. On The Subject of A Flood Propagation Computation Method (Muskingum Method). *J. Hydraul. Res.,* 7(2), pp. 205-230.

Dembélé, M. et al., 2020. Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns with Multiple Satellite Data Sets. *Water Resour. Res.*, 56(1), p. e2019WR026085.

Demirel, M. C. et al., 2018. Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrol. Earth Syst. Sci.*, 22(2), pp. 1299-1315.

DiMiceli, C. et al., 2015. *MOD44B MODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250m SIN Grid V006 [Data set]*. s.l.:NASA EOSDIS Land Processes DAAC.

Do, H. X., Gudmundsson, L., Leonard, M. & Westra, S., 2018. The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth Syst. Sci. Data*, 10(2), pp. 765-785.

Falcone, J. A., 2017. U.S. Geological Survey GAGES-II time series data from consistent sources of land use, water use, agriculture, timber activities, dam removals, and other historical anthropogenic influences. U.S. Geological Survey data release, 10.5066/F7HQ3XS4. Date accessed: 3/10/2022.

Fry, L. M. et al., 2014. The Great Lakes Runoff Intercomparison Project Phase 1: Lake Michigan (GRIP-M). *J. Hydrol.*, Volume 519, pp. 3448-3465.

Gardner, M. A. et al., 2018. Input data processing tools for the integrated hydrologic model GSFLOW. *Environ. Modell. Softw.*, Volume 109, pp. 41-53.

Ghiggi, G., Humphrey, V., Seneviratne, S. I. & Gudmundsson, L., 2019. GRUN: an observation-based global gridded runoff dataset from 1902 to 2014. *Hydrol. Earth Syst. Sci.*, 11(4), pp. 1655-1674.

Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, 377(1), pp. 80-91.

Hay, L. E. et al., 2006. Step wise, multiple objective calibration of a hydrologic model for a snowmelt dominated basin. *J. Am. Water Resour. As.*, 42(4), pp. 877-890.

He, X. et al., 2021. Climate-informed hydrologic modeling and policy typology to guide managed aquifer recharge. *Sci. Adv.*, 7(17), eabe6025.

Hengl, T. et al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2), p. e0169748.

Herman, M. R. et al., 2018. Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *J. Hydrol.*, Volume 556, pp. 39-49.

Hobeichi, S., Abramowitz, G., Evans, J. & E, B. H., 2019. Linear Optimal Runoff Aggregate (LORA): a global gridded synthesis runoff. *Hydrol. Earth Syst. Sci.*, 23(2), pp. 851-870.

Howarth, R. J., 1996. Sources for a History of the Ternary Diagram. *Brit. J. Hist. Sci.,* 29(3), pp. 337-356.

Huang, Q. et al., 2020. Using Remote Sensing Data-Based Hydrological Model Calibrations for Predicting Runoff in Ungauged or Poorly Gauged Catchments. *Water Resour. Res.,* 56(8), p. e2020WR028205.

Hunt, R. J., Feinstein, D. T., Pint, C. D., & Anderson, M. P., 2006. The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, northern Wisconsin. *J. Hydrol.,* 321(1-4), pp. 286-296.

Hunt, R. J. et al., 2013. *Simulation of Climate-Change Effects on Streamflow, Lake Water Budgets, and Stream Temperature Using GSFLOW and SNTEMP, Trout Lake Watershed, Wisconsin*, Reston, VA: U.S. Geological Survey.

Huntington, J. L. & Niswonger, R. G., 2012. Role of surface-water and groundwater interactions on projected summertime streamflow in snow dominated regions: An integrated modeling approach. *Water Resour. Res.,* 48(11), p. W11524.

Jensen, M. E., Robb, D. C. N. & Franzoy, C. E., 1970. Scheduling Irrigations Using Climate-Crop-Soil Data. *J. Irr. Drain. Div.*, 96(1), pp. 25-38.

Knoben, W. J. M. et al., 2020. A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. *Water Resour. Res.,* 56(9), p. e2019WR025975.

Kunnath-Poovakka, A., Ryu, D., Renzullo, L. J. & George, B., 2016. The efficacy of calibrating hydrologic model using remotely sensed evapotranspiration and soil moisture for streamflow prediction. *J. Hydrol.,* Volume 535, pp. 509-524.

Latifovic, R. et al., 2012. 2010 Land Cover of North America at 30 meters. In: *Remote Sensing of Land Use and Land Cover: Principles and Applications.* Boca Raton: CRC Press, pp. 303-321.

Lehner, B., Verdin, K. & Jarvis, A., 2008. New Global Hydrography Derived From Spaceborne Elevation Data. *Eos Trans. AGU,* 89(10), p. 93.

Livneh, B. & Lettenmaier, D. P., 2012. Multi-criteria parameter estimation for the Unified Land Model. *Hydrol. Earth Syst. Sci.,* 16(8), pp. 3029-3048.

Li, Y., Grimaldi, S., Pauwels, V. R. N. & Walker, J. P., 2018. Hydrologic model calibration using remotely sensed soil moisture and discharge measurements: The impact on predictions at gauged and ungauged locations. *J. Hydrol.,* Volume 557, pp. 897-909.

Liu, J., Liu, Q. & Yang, H., 2016. Assessing water scarcity by simultaneously considering environmental flow requirements, water quantity, and water quality. *Ecol. Indic.,* Volume 60, pp 434-441.

López López, P. L. et al., 2017. Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrol. Earth Syst. Sci.,* 21(6), pp. 3125-3144.

Luojus, K. et al., 2013. *Global snow monitoring for climate research: Algorithm Theoretical Basis Document (ATBD) – SWE-algorithm,* s.l.: European Space Agency.

Mai, J. et al., 2022. The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL). *Hydrol. Earth Syst. Sci.,* 26(13), pp. 3537-3572.

Markstrom, S. L. et al., 2008. GSFLOW-Coupled Ground-water and Surface-water FLOW model based on the integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005). In: U.S. Geological Survey *Techniques and Methods 6–D1*, 240p.

Markstrom, S. R. R. S. et al., 2015. PRMS-IV, the precipitation-runoff modeling system, version 4. In: U.S. Geological Survey *Techniques and Methods 6–B7*, 158p.

Martens, B. et al., 2017. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev.,* 10(5), pp. 1903-1925.

Matott, L. S., 2016. *OSTRICH–An Optimization Software Toolkit for Research Involving Computational Heuristics*, New York, USA: State University of New York at Buffalo.

McMillan, H. K., Westerberg, I. K. & Krueger, T., 2018. Hydrological data uncertainty and its implications. *WIREs Water,* 5(6), p. e1319.

Mei, Y., Reeves, H. & Mai, J., 2022. PRMS Model Archive for Selected Catchments in the Lake Michigan Basin Used in Examination of Multi-Objective Model Calibration: U.S. Geological Survey data release, 10.5066/P9DOVISZ. Date accessed: XXX.

Menne, M. J. et al., 2012. An Overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Ocean. Tech.,* 29(7), pp. 897-910.

Miralles, D. G. et al., 2011. Global land-surface evaporation estimated from satellite-based observations. *Hydrol. Earth Syst. Sci.,* 15(2), pp. 453-469.

Mostafaie, A., Forootan, E., Safari, A. & Schumacher, M., 2018. Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data. *Computat. Geosci.,* 22(3), pp. 789-814.

O, S. & Orth, R, 2021. Global soil moisture data derived through machine learning trained with in-situ measurements. *Sci. Data,* 8(1), p. 170.

Priestley, C. H. B. & Taylor, R. J., 1972. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Mon. Weather Rev.,* 100(2), pp. 81-92.

Rajib, A., Evenson, G. R., Golden, H. E. & Lane, C. R., 2018. Hydrologic model predictability improves with spatially explicit calibration using remotely sensed evapotranspiration and biophysical parameters. *J. Hydrol.,* Volume 567, pp. 668-683.

Rajib, M. A., Merwade, V. & Yu, Z., 2016. Multi-objective calibration of a hydrologic model using spatially distributed remotely sensed/in-situ soil moisture. *J. Hydrol.,* Volume 536, pp. 192-207.

Rakovec, O., Kumar, R., Attinger, S. & Samaniego, L., 2016. Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resour. Res.,* 52(10), pp. 7779-7792.

Regan, R. S. et al., 2019. The U. S. Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. *Environ. Modell. Softw.,* Volume 111, pp. 192-203.

Rientjes, T. H. M. et al., 2013. Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *J. Hydrol.,* Volume 505, pp. 276-290.

Thornton, M. M. et al., 2021. Gridded daily weather data for North America with comprehensive uncertainty quantification. *Sci. Data,* 8(1), pp. 190.

Thornton, P. E., Running, S. W. & White, M. A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.,* 190(3), pp. 214-251.

Tolson, B. A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resour. Res.,* 43(1).

Valente, F., David, J. S. & Gash, J. H. C., 1997. Modelling interception loss for two sparse eucalypt and pine forests in central Portugal using reformulated Rutter and Gash analytical models. *J. Hydrol.,* 190(1), pp. 141-162.

Van Loon, A. F., & Laaha, G., 2015. Hydrological drought severity explained by climate and catchment characteristics. *J. Hydrol.,* Volume 526, pp. 3-14.

Wolock, D. M., 2003. *Base-flow index grid for the conterminous United States.* U.S. Geological Survey data release, 10.3133/ofr03263. Date accessed: 10/1/2021.

Xie, K. et al., 2021. Identification of spatially distributed parameters of hydrological models using the dimension-adaptive key grid calibration strategy. *J. Hydrol.,* Volume 598, p. 125772.

Xu, S. et al., 2021. Investigating groundwater-lake interactions in the Laurentian Great Lakes with a fully-integrated surface water-groundwater model. *J. Hydrol.,* Volume 594, p. 125911.

Xu, T. et al., 2019. Evaluation of twelve evapotranspiration products from machine learning, remote sensing and land surface models over conterminous United States. *J. Hydrol.,* Volume 578, p. 124105

Yassin, F. et al., 2017. Enhanced identification of a hydrologic model using streamflow and satellite water storage data: A multicriteria sensitivity analysis and optimization approach. *Hydrol. Process.,* 31(19), pp. 3320-3333.