Uncertain Benefits of Using Remotely Sensed Evapotranspiration for Streamflow Estimation-Insights from a Randomized, Large-Sample Experiment

Tam Van Nguyen[1], Hung T.T Nguyen[2], Vinh Ngoc Tran[3], Manh-Hung Le[4], Binh Quang Nguyen[5], Hung Thanh Pham[6], Tu Hoang Le[7], Doan Van Binh[8], Thanh Duc Dang[9], Hoang Tran[10], and Hong Xuan Do[11]

[1]Helmholtz Centre for Environmental Research
[2]Columbia University
[3]University of Michigan
[4]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center
[5]The University of Danang - University of Science and Technology
[6]The University of Danang
[7]Research Center for Climate Change, Nong Lam University - Ho Chi Minh City
[8]Master Program in Water Technology, Reuse and Management, Vietnamese German University
[9]University of South Florida
[10]Pacific Northwest National Laboratory
[11]Nong Lam University

January 20, 2023

Abstract

Remotely sensed evapotranspiration (ET_{RS}) is increasingly used for streamflow estimation. Earlier reports are conflicting as to whether ET_{RS} is useful in improving streamflow estimation skills. We believe that it is because earlier works used calibrated models and explored only small subspaces of the complex relationship between model skills for streamflow (Q) and ET. To shed some light on this complex relationship, we design a novel randomized, large sample experiment to explore the full ET-Q skill space, using seven catchments in Vietnam and four global ET_{RS} products. For each catchment and each ET_{RS} product, we employ 10,000 SWAT (Soil and Water Assessment Tool) model runs whose parameters are randomly generated via Latin Hypercube sampling. We then assess the full joint distribution of streamflow and ET skills using all model simulations. Results show that the relationship between ET and streamflow skills varies with regions, ET_{RS} products, and the selected performance indices. This relationship even changes with different ranges of ET skills. Parameter sensitivity analysis indicates that the most sensitive parameters could have opposite contributions to ET and streamflow skills. Conditional probability assessment reveals that with certain ET_{RS} products, the probabilities of having good streamflow skills are high and increase with better ET skills, but for other ET_{RS} products, good model skills for streamflow are only achievable with certain intermediate ranges of ET skills, not the best ones. Overall, our study provides a useful approach for evaluating the value of ET_{RS} for streamflow estimation.

Hosted file

uncertain-benefits-of-using-remotely-sensed-evapotranspiration-for-streamflow-estimation-insights-from-a-randomized-large-sample-experiment

Uncertain Benefits of Using Remotely Sensed Evapotranspiration for Streamflow Estimation—Insights from a Randomized, Large-Sample Experiment

Tam V. Nguyen[1], Hung T.T. Nguyen[2,*], Vinh Ngoc Tran[3], Manh-Hung Le[4,5], Binh Quang Nguyen[6], Hung T. Pham[6], Tu Hoang Le[7], Doan Van Binh[8], Thanh Duc Dang[9], Hoang Tran[10], Hong Xuan Do[11,12*]

[1]Department of Hydrogeology, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany

[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA

[3]Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109, USA

[4]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

[5]Science Applications International Corporation, Greenbelt, MD 20771, USA

[6]The University of Danang - University of Science and Technology, Da Nang 550000, Vietnam

[7]Research Center for Climate Change, Nong Lam University – Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

[8]Master Program in Water Technology, Reuse and Management, Vietnamese German University, Ben Cat, Binh Duong 820000, Vietnam

[9]Department of Civil and Environmental Engineering, University of South Florida, Tampa, FL 33620, USA

[10]Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA 99354, USA

[11]Faculty of Environment and Natural Resources, Nong Lam University – Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

[12]Center for Technology Business Incubation, Nong Lam University – Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

*Corresponding authors: Hung T.T. Nguyen (hnguyen@ldeo.columbia.edu) and Hong X. Do (doxuanhong@hcmuaf.edu.vn)

Key points

- The relationship between model skills for streamflow and evapotranspiration is explored using a stochastic approach
- The value of remotely sensed evapotranspiration for streamflow estimation varies with regions, satellite products, and performance indices
- The probability of having good model skill for streamflow does not always increase with increasing model skill for evapotranspiration

Abstract

Remotely sensed evapotranspiration (ET_{RS}) is increasingly used for streamflow estimation. Earlier reports are conflicting as to whether ET_{RS} is useful in improving streamflow estimation skills. We believe that it is because earlier works used calibrated models and explored only small subspaces of the complex relationship between model skills for streamflow (Q) and ET. To shed some light on this complex relationship, we design a novel randomized, large sample experiment to explore the full ET-Q skill space, using seven catchments in Vietnam and four global ET_{RS} products. For each catchment and each ET_{RS} product, we employ 10,000 SWAT (Soil and Water Assessment Tool) model runs whose parameters are randomly generated via Latin Hypercube sampling. We then assess the full joint distribution of streamflow and ET skills using all model simulations. Results show that the relationship between ET and streamflow skills varies with regions, ET_{RS} products, and the selected performance indices. This relationship even changes with different ranges of ET skills. Parameter sensitivity analysis indicates that the most sensitive parameters could have opposite contributions to ET and streamflow skills. Conditional probability assessment reveals that with certain ET_{RS} products, the probabilities of having good streamflow skills are high and increase with better ET skills, but for other ET_{RS} products, good model skills for streamflow are only achievable with certain intermediate ranges of ET skills, not the best ones. Overall, our study provides a useful approach for evaluating the value of ET_{RS} for streamflow estimation.

Plain Language Summary

Evapotranspiration (ET), the amount of water evaporated from the Earth's surface through water bodies, soil, and plants, is an important component of the water cycle. It is often measured from space. These measurements are called remotely sensed ET (ET_{RS}) and are increasingly used to improve estimates of the water cycle. However, earlier studies reported conflicting results as to whether using ET_{RS} actually improves hydrological model performance. They calibrated their models with and without ET_{RS} to see whether including ET_{RS} would help simulating streamflow (river discharge), and found that it did in some cases but did not in other cases. To understand the added value of ET_{RS} in model calibration, we design a novel experiment that is counter-intuitive at first sight: we do not calibrate our models; instead, we test 10,000 random models to see the full range of their performance—how well they simulate streamflow in relation to how well they simulate ET. We show that the relationship between ET and streamflow performance is complex, and the value of using ET_{RS} for streamflow estimation is uncertain as it depends on where the calibrated models land on this space.

1 Introduction

In recent decades, advances in remote sensing have facilitated the application of hydrological models—areas lacking ground observations may now be compensated by remotely sensed data (Dile et al., 2020). Remote sensing products have provided information on different components of the terrestrial water cycle at various spatial and temporal resolutions, for example, precipitation (Hsu et al., 1997), evapotranspiration (Mu et al., 2013; Senay et al., 2013), soil moisture (Hornáček et al., 2012), groundwater storage dynamics (Tapley et al., 2004), lake water levels (Crétaux et al., 2011), and snow cover (Hall et al., 1995, 2002; Tran et al., 2019). Remote sensing products have been used in addition to ground observations as model inputs since they can provide better spatiotemporal coverages (Baez-Villanueva et al., 2020; Liu et al., 2017). In ungauged or poorly gauged catchments, remote sensing products have been demonstrated as a potential source of data for streamflow estimation (Huang et al., 2020; Kunnath-Poovakka et al., 2016; Zhang et al., 2020).

Evapotranspiration (ET) is an important component of the hydrological cycle—about 60% of the Earth's terrestrial precipitation returns to the atmosphere as evapotranspiration (Pan et al., 2015; Trenberth et al., 2009). ET-related variables have been extensively observed from space. Several remotely sensed ET (ET_{RS}) products are available at the global scale with long temporal (decadal) coverage (Mu et al., 2013; Senay et al., 2013). In recent years, ET_{RS} products have been increasingly used by the hydrological modeling community, as model input or as calibration data (Herman et al., 2018; Immerzeel & Droogers, 2008; Kunnath-Poovakka et al., 2016; Zhang et al., 2009). Taking advantage of ET_{RS} products with their global coverage is a promising approach to improve streamflow estimate (Martens et al., 2017; Mu et al., 2013). Evaluating the value of ET_{RS} for streamflow estimation is especially important considering that a majority of the world's river reaches do not have stream gauges installed to monitor flow (Krabbenhoft et al., 2022).

Among pioneering works that evaluated the value of ET_{RS} for streamflow estimation, Immerzeel & Droogers (2008) calibrated SWAT (Soil and Water Assessment Tool, Arnold et al., 1998) models against Moderate Resolution Imaging Spectroradiometer (MODIS) derived ET for the Upper Bhima catchment (India). Their results showed that, qualitatively, the calibrated model is better at producing streamflow that resembled observations relative to the uncalibrated one. Later works quantified model performance for streamflow and ET under different calibration schemes, and results were inconclusive. For example, Zhang et al. (2009) calibrated a simple lumped model against (i) streamflow only, and (ii) both streamflow and ET_{RS}. They found that the former had better performance for streamflow compared to the latter, suggesting that adding ET_{RS} to the calibration process was not helpful. Herman et al. (2018) found that calibrating SWAT models against ET_{RS} significantly reduced streamflow estimation skills, while a multi-objective calibration scheme targeting both streamflow and ET improved the model performance for ET while maintaining an acceptable level of skills for streamflow. Nguyen et al. (2020a) found that the use of MODIS-derived ET does not affect model performance for streamflow since model performance for ET and streamflow was highly positively correlated (only for behavioral simulations for Q and ET). Many other studies (Dembélé et al., 2020; Demirel et al., 2018; Gui et al., 2019; Jiang et al., 2020; Kunnath-Poovakka et al., 2016; Parajuli et al., 2018; Rajib et al., 2018; Sirisena et al., 2020; Willem Vervoort et al., 2014; Zhang et al., 2020) using various ET_{RS} products and a wide range of models and calibration techniques, came to different conclusions (see Table S1). In summary, various experiments with numerous setups found that the value of having ET_{RS} ranges from positive, neutral, to negative.

This paradox suggests that the relationship between ET and streamflow skills is complex: there is sometimes a trade-off between ET skill (model performance for ET) and streamflow skill (model performance for streamflow) but not always. One common feature among previous experiments is that all of them calibrated models and evaluated model skills upon validation. We contend that using only a small set of calibrated models is insufficient to explore the complex relationship between ET and streamflow simulation skills. This is because different calibration schemes navigate towards different subspaces of the streamflow-ET skill relationship, leading to different conclusions.

To shed some light on this complex relationship, we design a randomized, large sample experiment. Instead of calibrating hydrological models with and without ET_{RS} and evaluating model performance post-calibration, as prior studies did, we simply generate a large number of models with random parameter values and calculate their skill scores with respect to ET and streamflow. Our approach may seem counter-intuitive at first, but there are two reasons that merits randomization over calibration. First, we can examine the full ET-streamflow skill space instead of a few points in that space from some calibrated models. The second reason lies in the randomness nature of model skills. Semi-distributed and distributed models are complex and thus prone to overparameterization (Beven, 2006)—i.e., models may be overfitted to small training data size—a problem particularly pertinent to poorly gauged basins. Thus, even after a model is calibrated, there is little guarantee that model skills are robust during validation or regionalization. In other words, model skills during validation and regionalization are essentially random.

To demonstrate this approach, we use four global ET_{RS} products and seven catchments in Vietnam (with diverse catchment characteristics and contrasting ET and streamflow regimes). For each catchment–ET_{RS} pair, we simulate 10,000 SWAT models with randomized parameters to obtain a large ensemble of simulated streamflow and ET. We then use conditional probability to assess how likely a model is good for ET is good for streamflow simulation and *vice versa*. While our study is limited to specific regions, ET_{RS} products, and a hydrological model, our findings could provide a useful approach for evaluating the value of ET_{RS} for streamflow estimation in the study area and beyond.

2 Study Area and Data

2.1 Study Area

We selected seven catchments across Vietnam (Figure 1) to evaluate the use of ET_{RS} products for streamflow modeling. These catchments do not have large dams, large urban areas, or substantial changes in land use during the 2000-2019 periods (Do et al., 2022). They cover a wide range of attributes, for example, catchment area ranges from 603 to 6392 km^2, and areal percentages of forest land range from 6.2 to 84.9% (Table 1). The selected catchments are located in both lowland (median elevation of 106.5 m above mean sea level – m.a.s.l) and mountainous (median elevation of 1406 m.a.s.l) areas. The selected catchments represent seven Vietnamese sub-climatological regions (D. N. Nguyen & Nguyen, 2004; Phan et al., 2009). The four catchments in Central and Southern Vietnam (GSO, CDA, SDI and AHO) receive more annual rainfall than do the catchments in Northern Vietnam (CHU, XLA, and NKH). The runoff coefficients of SDI and AHO catchments (0.90 and 0.82, respectively) are significantly higher than those of the other catchments, indicating that evaporative losses are quite small in these catchments compared to the others.

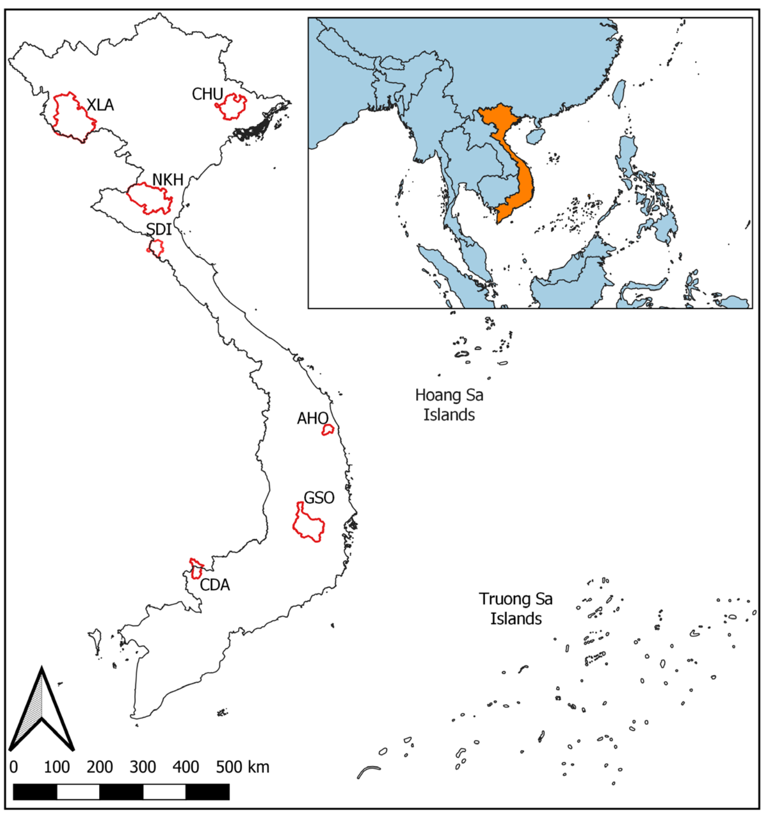

Figure 1. Location of the seven study catchments in Vietnam. The short names CHU, XLA, NKH, GSO, CDA, SDI, and AHO stand for Chu, Xa La, Nghia Khanh, Giang Son, Can Dang, Son Diem, and An Hoa catchments, respectively (Do et al., 2022).

Table 1. Characteristics of the seven study catchments.

Catchment ID	CHU	XLA	NKH	GSO	CDA	SDI	AHO
Area (km^2)	2176	6449	4315	3181	752	827	392
Runoff depth[a] (mm/yr)	600.4	592.1	840.9	723.6	600.9	1629.7	2660.4
Precipitation[a] (mm/yr)	1555.3	1479.1	1558.4	1802.0	1913.3	1994.5	2948.6
Runoff coefficient	0.39	0.40	0.54	0.40	0.31	0.82	0.90
Temperature[a] (°C)	22	22.2	25.3	24.2	28.4	24.4	24.9
Forest[b] (%)	31.7	36	48.3	42.2	6.2	84.9	81.5
Agriculture[b] (%)	68.1	63.9	51.5	56.8	93.8	15.1	18.5
Elevation[c] (m.a.s.l)	502.5	1190	1211.5	1406	106.5	956.5	512.5
Catchment slope[c] (%)	25.8	39.4	26.7	14.3	10.22	33.7	32.7

[a]mean annual value from 2010 to 2019, [b]areal percentage, [c]median value

2.2 Input data for SWAT

We used the Soil and Water Assessment Tool (SWAT), a semi-distributed hydrological model that has been used widely in water research, to support our investigation (Arnold et al., 1998, 2012). Data for several SWAT input variables, including Digital Elevation Model (DEM), land use, soil, and weather, were collected. A 30 m spatial resolution DEM product (ASTER, Advanced Spaceborne Thermal Emission and Reflection) released by the National Aeronautics and Space Administration (NASA) in collaboration with Japan's Ministry of Economic, Trade, and Industry, was downloaded from the USGS Earth Explorer website (https://earthexplorer.usgs.gov/). Land use data were obtained from the European Space Agency Climate Change Initiative Land Cover data set (ESA-LC, https://www.esa-landcover-cci.org/), which provides global land cover maps at 300 m spatial resolution between 1992–2019. This data set has been validated in several regions in Asia and Africa, demonstrating its good agreement with ground observation (ESA, 2017). Here we use the ESA-LC data set in the year 2000. In addition, soil data were obtained from the Harmonized World Soil Database (HWSD) version 1.2 (Fischer et al., 2008). HWSD is a 30 arc-second raster database with over 15,000 different soil mapping units that combine existing regional and national updates of soil information. Daily streamflow observations at the catchment outlets from 2010 to 2019 were obtained from the Vietnam Meteorological and Hydrological Administration. For climate data, daily precipitation was collected from local meteorological stations in each river basin, daily maximum and minimum air temperature data, solar radiation, relative humidity, and wind speed data were collected from the Global Land Data Assimilation System (https://ldas.gsfc.nasa.gov/data; Rodell et al., 2004) for the period 2010–2019.

2.3 Remote Sensing Evapotranspiration Products

We used four global ET_{RS} products (actual ET), namely, (1) the Global Land Evaporation Amsterdam Model (GLEAM, Martens et al., 2017), (2) the Moderate Resolution Imaging Spectroradiometer (MOD16A2; Mu et al., 2013), (3) the operational Simplified Surface Energy Balance model (SSEBop; Senay et al., 2013), and (4) TerraClimate (Abatzoglou et al., 2018). These ET_{RS} products are available at different spatiotemporal resolutions and are derived using different input data and techniques (Table 2). GLEAM and MOD16A2 use only satellite-based data to estimate ET. SSEBop uses both satellite observations and ground-based weather data as model input, while TerraClimate depends mainly on ground-based measurements. Three models (MOD16A2, SSEBop, and TerraClimate) are based on the Penman-Monteith (P-M) (Allen, 1986; Monteith, 1965) equation to estimate reference potential ET, while GLEAM is based on the Priestley-Taylor (P-T, Priestley & Taylor, 1972) equation, which is a simplified solution of the P-M equation. The daily GLEAM ET product and the 8-day MOD16A2 ET product were aggregated to the monthly time step. ET_{RS} data sets were spatially and temporally to catchment-scale and monthly time step, respectively, for evaluating with SWAT outputs.

Table 2. List of the four ET_{RS} products used in this study.

ET_{RS} products	Spatial/ temporal resolution	Potential ET method	Spectral/field measurements
GLEAM	25 km/daily	Priestley-Taylor	Red, NIR, PMW, AMW
MOD16A2	0.5 km/8-day	Penman–Monteith	Red, NIR
SSEBop	1 km/monthly	Penman–Monteith	Red, NIR, TIR, NOAA GDAS
TerraClimate	4 km/monthly	Penman–Monteith	WorldClim, CRU, JRA-55

NIR = Near InfraRed; TIR = Thermal InfraRed; PMW = Passive Microwave; AMW = Active Microwave; NOAA GDAS = National Oceanic and Atmospheric Administration Global Data Assimilation System; CRU = Climate Research Unit; JRA = Japanese 55-year Reanalysis

The time series of the four ET_{RS} products in each catchment are shown in Figure 2. In CHU, XLA, NKH, and SDI, the four ET_{RS} products generally agree with one another. There are large discrepancies among the products at GSO, and, to a lesser extent, CDA and AHO, showing the spatial and temporal uncertainties among these products.

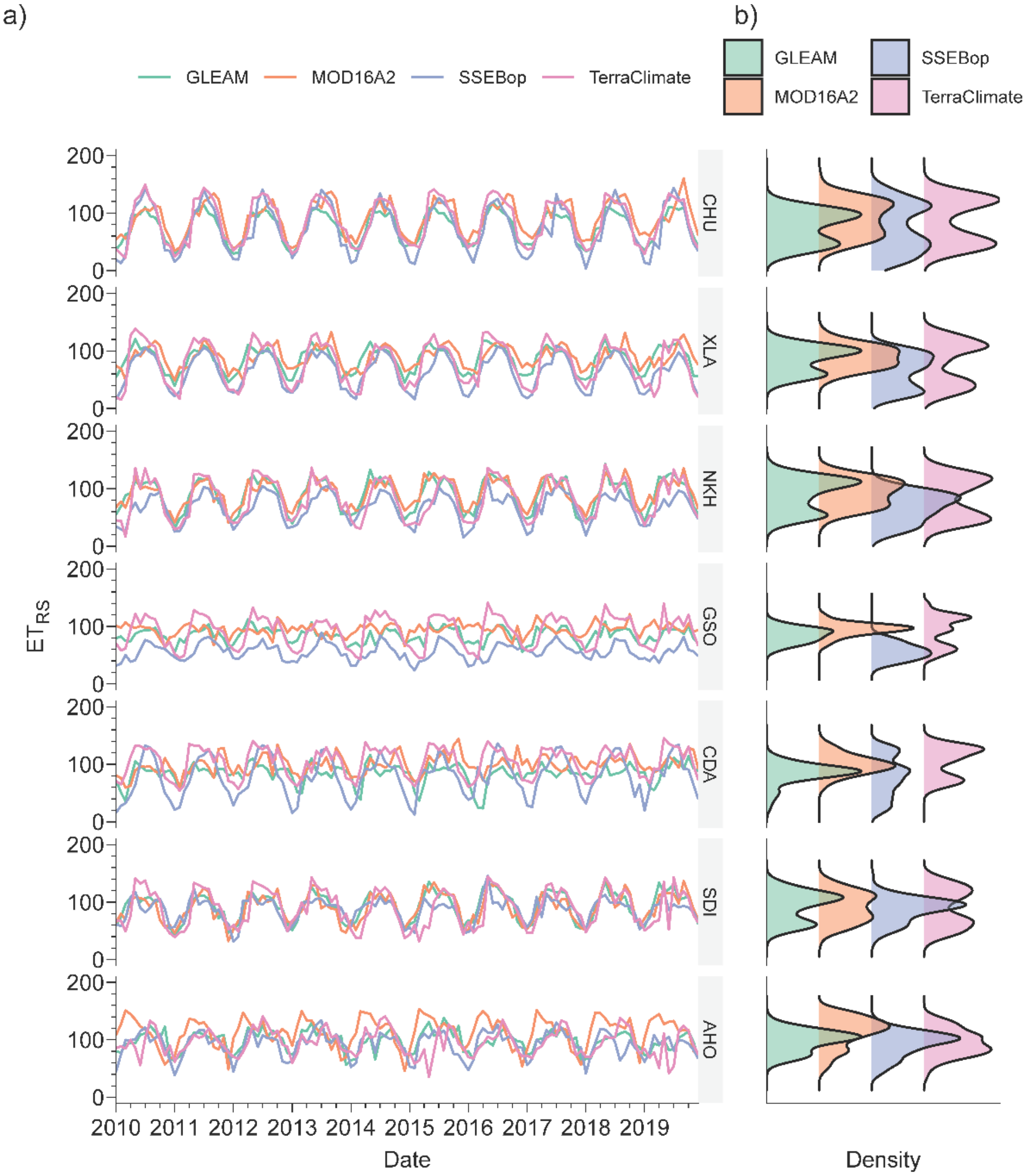

Figure 2. Temporal variation (a) and probability density function (b) of ET_{RS} from different products at each catchment.

3 Methodology

This work involves four main stages: simulation, skill distribution analysis, sensitivity analysis, and probabilistic assessment (Figure 3). In stage 1, we aim to produce a wide range of model skills for streamflow (Q) and ET. Therefore, for each catchment–ET_{RS} pair, we run 10,000 SWAT models, each of which has a different, randomized set of parameters. The model configuration and parameter randomization scheme are presented in Sections 3.1 and 3.2. This step yields 70,000 pairs of ET and streamflow time series (seven catchments with 10,000 model runs for each catchment). In stage 2, we calculated the goodness-of-fit of each simulated time series against its corresponding ET_{RS} products and observed streamflow, resulting in 280,000 pairs of ET and streamflow skill values (seven catchments, 10,000 model runs for each catchment, four ET_{RS} products). We also collected the best 100 NSE values for each case to understand the relationship between ET-streamflow skills in good models. In stage 3, sensitivity analysis was used to evaluate the effects of the most sensitive parameters (for both ET and streamflow) on the relationship between ET and streamflow skills (Section 3.4). Finally, in stage 4, the conditional probability of ET skill on a given range of streamflow skills was calculated to find which ET_{RS} products can produce better performances (Section 3.5), giving a statistical sense about the applicability of ET_{RS} in streamflow estimation. In the remainder of this section, we describe each step in detail.

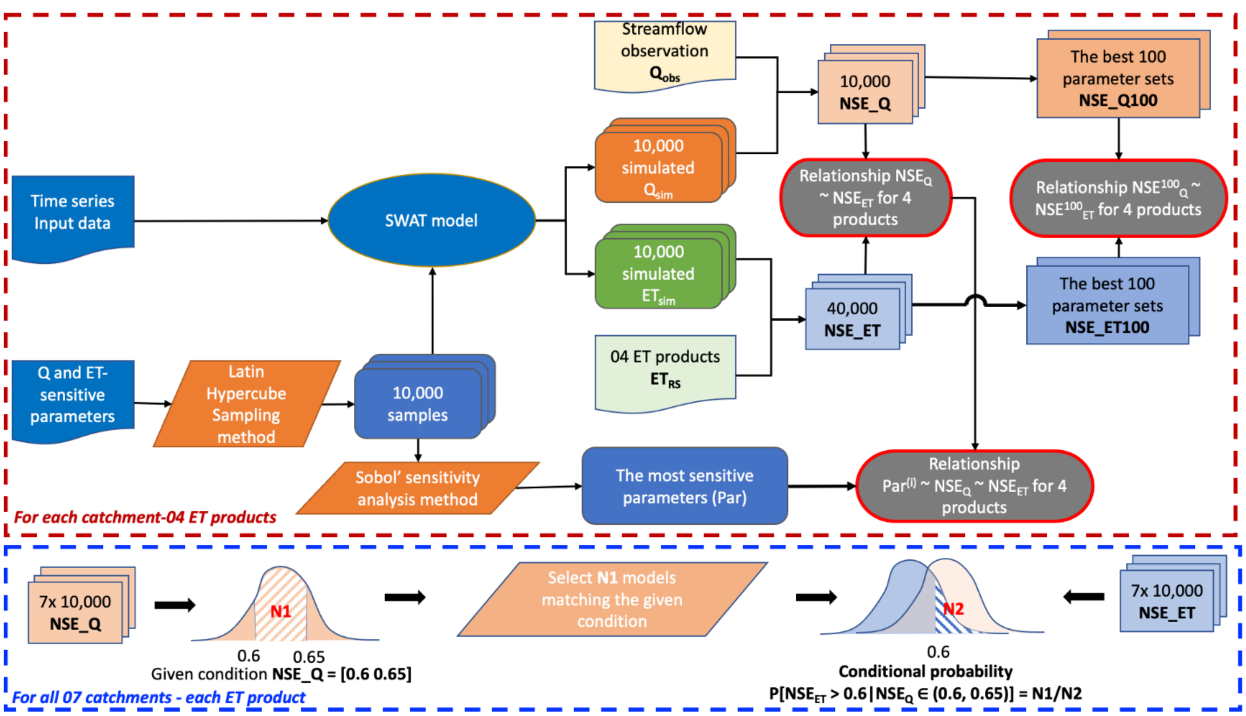

Figure 3. Flow chart of the research methodology employed in this study.

3.1 SWAT Model and Model Setup

In SWAT, a catchment is divided into subcatchments, which are further divided into Hydrologic Response Units (HRUs) (Neitsch et al., 2011). An HRU is an area of land within a subcatchment with a unique combination of land use, soil type, and topographic slope. SWAT simulates different phases of the water cycle, e.g., evapotranspiration, soil-water dynamics, groundwater flow, and streamflow. Actual evapotranspiration (hereafter referred to as ET) was then calculated based on potential ET following one of the available approaches: the Penman–Monteith (Allen, 1986; Allen et al., 1989; Monteith, 1965), Priestly–Taylor (Priestley & Taylor, 1972), and Hargreaves (Hargreaves & Samani, 1985), depending on data availability. A detailed description of the implementation of these approaches was described in the SWAT model documentation (Neitsch et al., 2011).

All of our SWAT models were set up using common settings. Specifically, (1) we used the same criteria for HRU definitions, (2) all models used the Penman–Monteith approach for calculating potential ET, and (3) all models were set to run at the daily time step from 2008-2009 with three years of warm-up (2008-2009) and ten years (2010-2019) for model ET- and Q-skill evaluation.

3.2 Parameter Randomization

Our goal is to generate a wide range of model skills with respect to both streamflow and ET. Therefore, instead of calibrating our models against streamflow and/or ET, we generated 10,000 random parameter sets for each catchment using the random Latin Hypercube Sampling (LHS) approach. The parameters and their ranges (Table 3) were selected based on our literature review of the most frequently used parameters for either ET or streamflow calibration (Neitsch et al., 2011; Nguyen et al., 2022a; Nguyen et al., 2020; Tobin & Bennett, 2017; Odusanya et al., 2019; Le et al., 2022). Including both ET- and streamflow-sensitive parameters allowed us to explore the uncertainty in streamflow when the models are calibrated for ET, and vice versa. We also conduct a sensitivity analysis after the models are simulated (Section 3.3). Parameter randomization and model execution were done in the R environment (R Core Team, 2021) with *R-SWAT* (Nguyen et al., 2022b).

Table 3. The selected parameters for randomization and their ranges. These parameter ranges were used for all catchments. The prefixes "r" and "v" indicate relative change and actual value, respectively.

Parameter	Description	Min	Max
r_CN2	SCS curve number II value (-)	-0.25	0.25
r_SOL_K	Soil saturated hydraulic conductivity (mm/hr)	-0.25	0.25
r_SOL_AWC	Soil available water content	-0.25	0.25
v_GW_DELAY	Groundwater delay (days)	10	500
v_ALPHA_BF	Baseflow alpha factor (days)	0	1
v_SHALLST	Initial depth of water in the shallow aquifer (mm)	0	1000
v_DEEPST	Initial depth of water in the deep aquifer (mm)	0	1000
v_GWQMN	Threshold baseflow to occur (mm)	0	1000
v_GW_REVAP	Threshold for groundwater 'revap' to occur (mm)	0.02	0.2
v_ESCO	Soil evaporation compensation factor (-)	0.01	1
v_EPCO	Plant uptake compensation factor (-)	0.01	1
v_CANMX	Maximum canopy storage (mm)	1	10
v_OV_N	Manning's "n" value for overland flow	0.01	0.3
v_CH_K2	Effective hydraulic conductivity (mm/hr)	0	25
v_CH_N2	Manning's n value for main channel	0.025	0.065
v_SURLAG	Surface runoff lag time (days)	0.1	0

3.3 Evaluation Metrics

For each catchment, we analyzed the relationship between the model skills for ET and streamflow using ET_{RS} products (Section 2.3) and observed streamflow. We used two common metrics: Nash-Sutcliffe Efficiency (NSE, Nash & Sutcliffe, 1970) and Kling-Gupta Efficiency (KGE, Gupta et al., 2009), to evaluate the model skills. In the main analysis, we will focus on the

NSE, and we provide additional results with the KGE in the Supplementary Information (see Section 4). The NSE is formulated as

$$NSE = 1 - \frac{\sum_{i=1}^{n}(x_i^{sim} - x_i^{obs})^2}{\sum_{i=1}^{n}(x_i^{sim} - \underline{x})^2} \quad (1)$$

where x^{sim} and x^{obs} are the simulated (from SWAT) and observed/reference values, respectively, $\underline{x}$ is the mean of the observations/reference values, and n is the number of observations/reference values.

We first calculated NSE for ET (NSE_{ET}) for all 10,000 simulated ET time series in each catchment against each ET_{RS} product. This step results in 40,000 NSE_{ET} values. We then calculated NSE for streamflow (NSE_Q) for all 10,000 simulated streamflow time series in each catchment against the respective observed streamflow time series. Finally, we explored the relationships between NSE_{ET} and NSE_Q for all 10,000 parameter sets in each catchment, as well as for the best 100 parameter sets in Q and the best 100 parameter sets in ET. The procedure is repeated for the KGE to assess the robustness of our findings.

3.4 Sensitivity Analysis

To understand how ET- and Q-sensitive parameters affect the model ET- and Q-skills, we first determined the most sensitive parameters for both ET and Q and then explored the relationships between the values of these parameters and skill scores. Sobol' sensitivity analysis (SA) was employed to identify key parameters and characterize parameter sensitivities (Saltelli, 2002; Sobol, 2001) as follows. First, using Analysis of Variance (ANOVA), the total variance of the NSE (or KGE) is decomposed into the variance contributions of individual parameters (Equation 2).

$$D(NSE\ or\ KGE) = \sum_{i=1}^{N} D_i + \sum_{j<i} D_{ij} + \ldots + D_{1\ldots N} \quad (2)$$

where D_i is the variance for the change of the *i*th model parameter, *N* the number of model parameters, D_{ij} the variance of the pairwise interaction of *i*th and *j*th parameters (two-way interactions), and $D_{1\ldots N}$ the N-way interaction term. An overall Sobol' sensitivity index is then determined for each parameter (Equation 3):

$$S_i(NSE) = 1 - \frac{D_{\underline{a}}}{D(NSE)} \quad (3)$$

where S_i is the main sensitivity index for the change of a parameter *i*, $D_{\underline{i}}$ is the variance averaged over the contributions resulting from all other parameters except *i*.

3.5 Assessment of model skills for each ET product using conditional probability

After the distribution of model skills is obtained, we assessed the probability that a model that is good for ET is also good for streamflow, and *vice versa*. We used a threshold of 0.6 for the NSE score to represent a good performance of a model for a variable (ET or streamflow). This threshold choice is somewhat arbitrary, but it is in line with the literature (Moriasi et al., 2007). Based on this threshold, we calculated the conditional probability that a model will have a good streamflow score given that it is within a certain ET score, as well as the conditional probability that a model will have a good ET score given that it is within a certain range of streamflow score. These probabilities were calculated separately for each ET product so as to evaluate these products, but were calculated over the catchments altogether (i.e., the total number of 70,000 models for each ET_{RS} product), as we aimed to generalize our findings for a "generic" unknown catchment. For example, the conditional probability $P[NSE_{ET} > 0.6|NSE_{Q} \in (0.6, 0.65)]$ for GLEAM is calculated as follows:

- Count all models whose NSE_Q is within (0.6, 0.65) across all catchments; this gives a number N_1.
- Count among N_1 the number of models whose NSE_{ET} with respect to GLEAM is above 0.6; this gives a number N_2.
- The ratio N_2/N_1 is then the desired probability.

The probability was then assessed to understand the complex relationship between Q- and ET-performance. This procedure was also repeated for the KGE to assess whether the findings vary substantially when different evaluation metrics are used.

4 Results and Discussion

4.1 Model Skills for ET and Streamflow

We first explore the relationship between model skills for ET and that for streamflow over each catchment (i.e., from 10,000 simulations for each ET_{RS} product). Figure 4 shows the results using NSE, in which two patterns of relationship between NSE_{ET} and NSE_Q are observed, and these patterns are similar across the four ET_{RS} products (Figure 4a). For five catchments (CHU, XLA, NKH, GSO, and CDA), we observe first a positive correlation between NSE_{ET} and NSE_Q, meaning that increased skill for ET is associated with an increased skill for Q. However, this is only true for the lower values of NSE, particularly with negative NSE_{ET}. As NSE_{ET} increases towards the highest ranges in each case, the positive correlation diminishes. It means that improving model skills for ET will not necessarily lead to an improvement in model skills for Q. Interestingly, a special case is observed in the GSO catchment with the SSEBop product, where NSE_{ET} correlates negatively with NSE_Q ($r = -0.73$, $p < 0.001$). This is the only case with a statistically significant negative correlation. On the other hand, we observe no clear relationships between NSE_{ET} and NSE_Q for the SDI and AHO catchments, where model skills tend to concentrate along two lines: a horizontal line with fairly similar NSE_Q, and a vertical line with fairly similar NSE_{ET}. Among the four ET_{RS} products, two satellite-based products (GLEAM and MOD16A2) generally resulted in lower skills for ET compared to partially and mainly ground-based products (SSEBop and TerraClimate).

From the 10,000 models, we selected those that are either in the best 100 models for NSE_Q or the best 100 models for NSE_{ET} (Figure 4b). Here, the trade-off between streamflow and ET prediction skills becomes apparent: the selected models lie along two perpendicular lines, closely resembling a Pareto frontier. In each catchment-product pair, the intersection of the best 100 models for streamflow and the best 100 models for ET consists of only 2–11 models. This means most models either produce high NSE_Q or high NSE_{ET}, and very few models could capture both processes. Positive NSE_Q was achieved for all catchments while NSE_{ET} was comparatively lower (often negative) and varied in a wider range across different ET_{RS} products, even within the same catchments and products (Figure 4b). This is due to the high uncertainties in different ET_{RS} products as also illustrated in Section 2.3 (Figure 2). The low skills even for the best models mean that it is difficult for SWAT models to capture ET as expressed in the ET_{RS} products in these tropical catchments. The reasons could be that SWAT is not suitable for these tropical catchments, or that the ET_{RS} products have limitations in this region, or both.

Results for the KGE metric (see Figure S1) show that the relationship between model performance for ET and streamflow also depends on the metric used. For example, with the GSO catchment and MOD16A2 product, a negative correlation between KGE_{ET} and KGE_Q (Figure S1a) is observed while that between NSE_{ET} and NSE_Q is positive (Figure 4a). It means that depending on a certain aspect of streamflow (reflect by the evaluation metric) the modelers are focusing on, ET_{RS} product could be useful or even have negative consequences for streamflow estimation. For example, the best 100 models for ET, in this case, have much lower KGE_Q compared to other KGE_Q from the models which have lower KGE_{ET} (Figure S1b, GSO catchment, MOD16A2 product). Furthermore, considering the uncertainty in ET_{RS} products, the use of ET_{RS} products for stream estimation in this case (negative correlation between KGE_{ET} and KGE_Q) is in question.

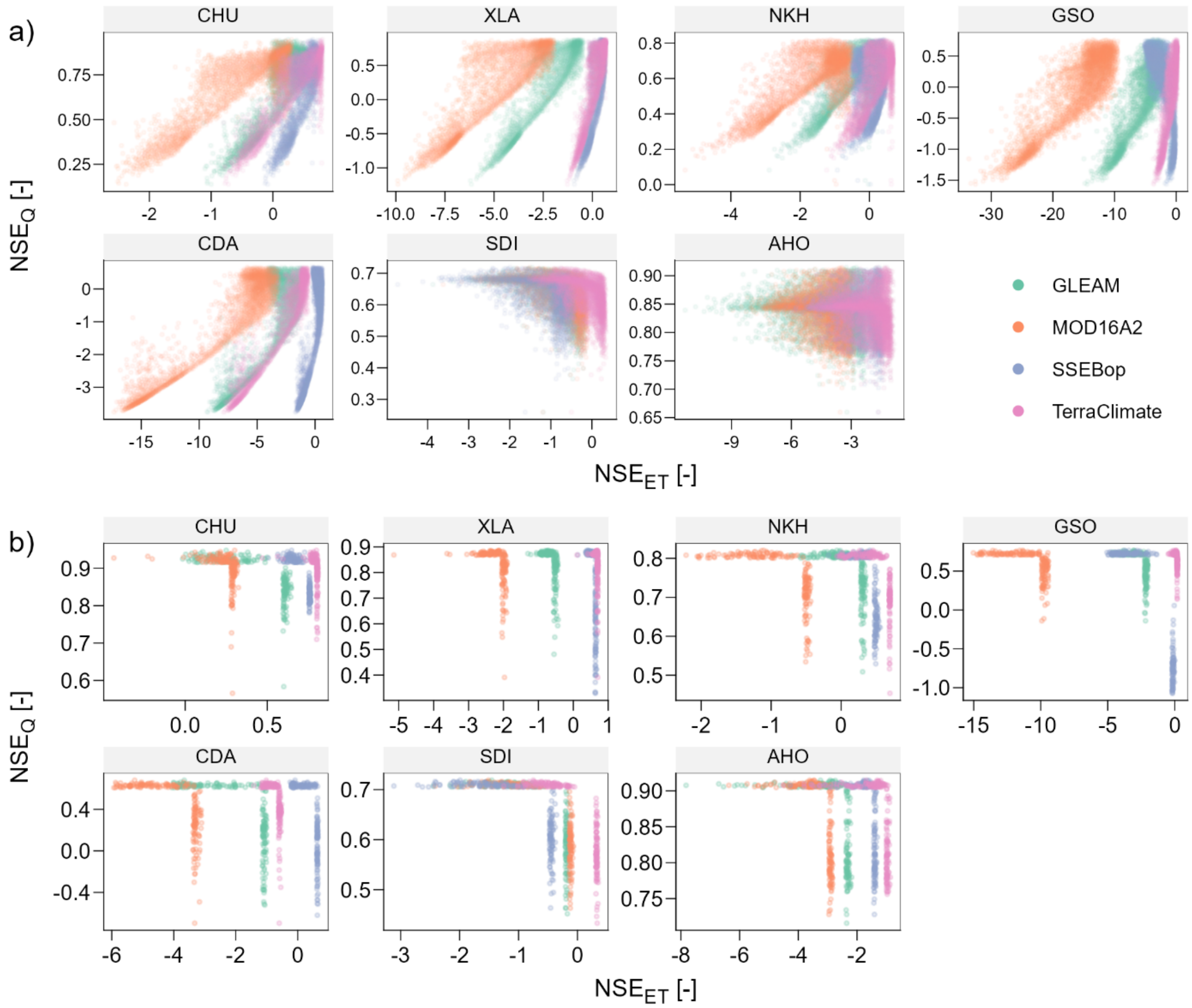

Figure 4. Distribution of NSE scores for ET (NSE_{ET}) versus NSE scores for streamflow (NSE_Q) for each catchment and ET_{RS} product. Panel **a** shows the scores of all 10,000 models and panel **b** shows the scores of models that are in either the top 100 for NSE_Q or the top 100 for NSE_{ET}. Note the large differences in x- and y-axis scales among the catchments.

4.2 Parameter Sensitivity

Figure 5 shows the total sensitivity of each parameter with respect to streamflow and ET (the objective function is $NSE_{ET} + NSE_Q$). In line with prior studies (e.g., Nguyen et al., 2020; Odusanya et al., 2019), we found that both streamflow and ET are highly sensitive to the curve number (CN2). In addition, ET is sensitive to soil evaporation compensation factor ESCO, and to a lesser extent, to soil available water content SOL_AWC. On the other hand, streamflow (Q) is sensitive to groundwater delay GW_DELAY and threshold to baseflow occur GWQMN, although the sensitivity varies among catchments. Results from the sensitivity analysis with the objective function is the KGE ($KGE_{ET} + KGE_Q$) show similar results in term of sensitivity ranking (e.g., both CN2 and ESCO are the most sensitive parameters among all catchments and ET_{RS} products), however, higher variation in the sensitive indices among different ETRS product (Figure S2). In the remaining, only results from the sensitivity analysis with the NSE as objective functions are shown.

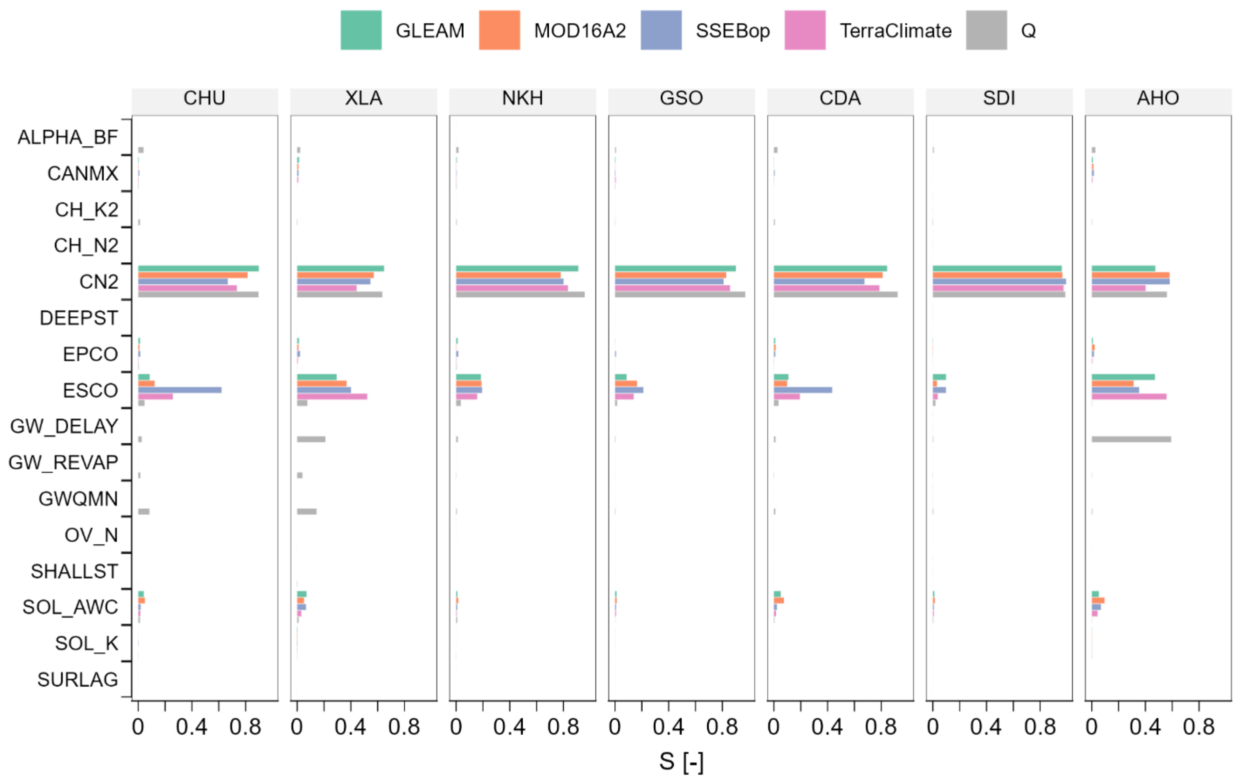

Figure 5. Total sensitivity (S) of streamflow and ET with respect to each model parameter in each catchment and variable (ET_{RS} product and observed streamflow Q).

Based on the results of parameter sensitivity analysis, we selected four parameters, namely CN2, ESCO, GW_DELAY, and GWQMN for further analysis. Figure 6 shows the relationships between the values of these four parameters and their NSE scores. As expected from the sensitivity analysis, NSE_{Q} and NSE_{ET} are strongly dependent on CN2, and two patterns can be observed. For the first group of five catchments (CHU, XLA, NKH, GSO, and CDA), the CN2–NSE_{ET} and CN2–NSE_{Q} relationships vary in the same direction: for both streamflow and ET, high values of CN2 are associated with low NSE, and NSE increases as CN2 decreases, to a certain threshold when NSE is much less or no longer dependent on CN2. This explains our observations in Figure 4a. At first, NSE_{ET} and NSE_{Q} increase together because they covary with CN2, and then in the higher NSE ranges, NSE_{ET} and NSE_{Q} no longer correlate with each other because they are less or no longer dependent on CN2.

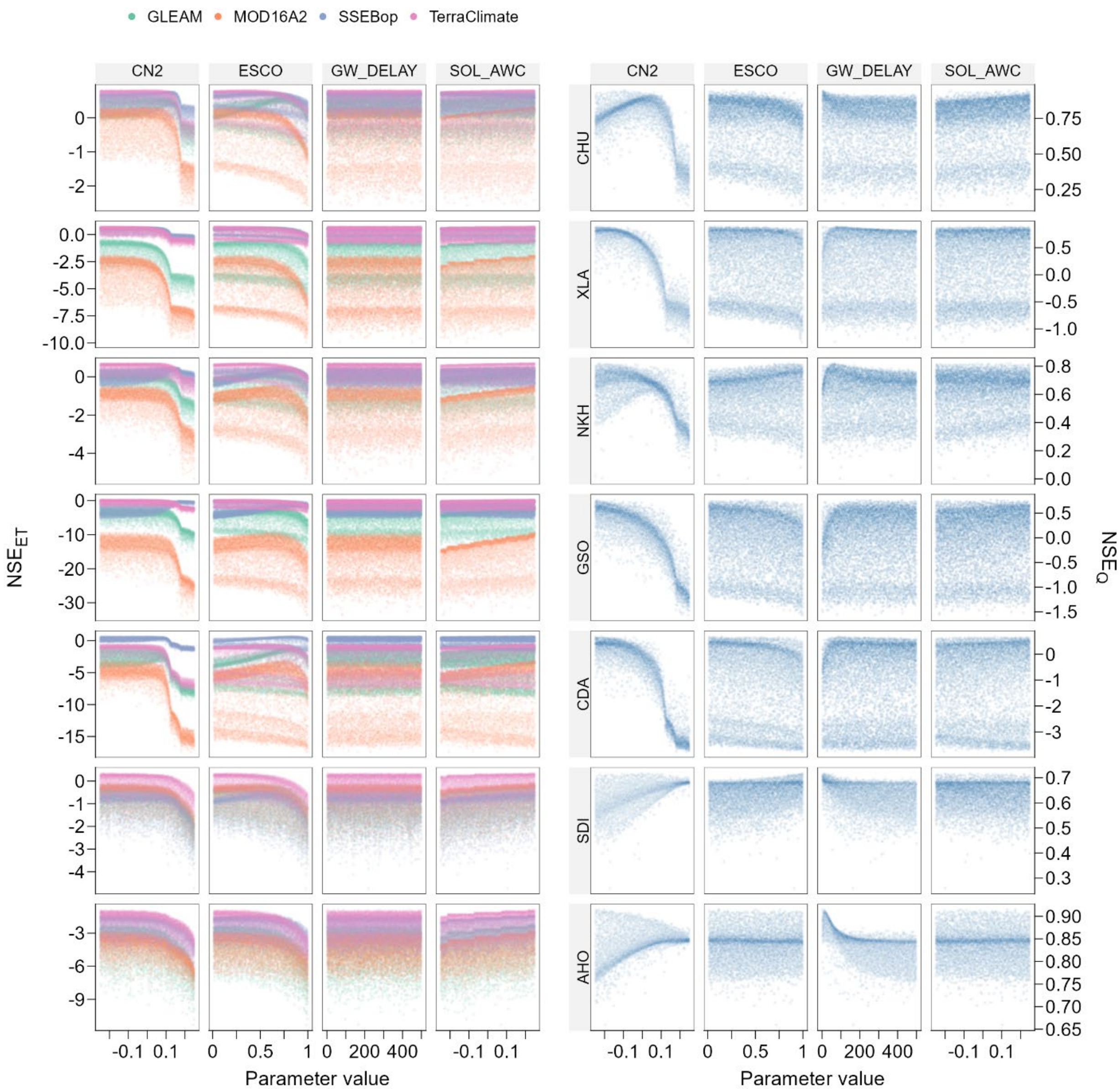

Figure 6. Relationships between model skills and parameter values for ET (first four columns) and streamflow (last four columns). Each row represents one catchment.

For the second group of catchments (AHO and SDI), the CN2–NSE_{ET} and CN2–NSE_{Q} relationships vary in *opposite* directions: high CN2 values are associated with low NSE_{ET} but high NSEQ, and vice versa. Again, this could explain the NSE_{Q}–NSE_{ET} relationship we observed for these two catchments in Figure 4a. As CN2 has opposite effects on NSE_{Q} and NSE_{ET}, models tend to concentrate on two perpendicular lines, one with high NSE_{ET} and low NSE_{Q}, and one with high NSE_{Q} and low NSE_{ET}.

Interestingly, in the region of high NSE where CN2 becomes less sensitive, some other parameters become more sensitive, although their sensitivity levels are less consistent across all catchments and products compared to that of CN2. For example, high NSE_Q values are sensitive to GW_DELAY, particularly in the AHO catchment (Figure 6, column 7). This means that model parameters do not have the same sensitivity throughout their ranges, and the relative sensitivity among parameters also changes. Therefore, it is important to explore a wide range of model skills and parameters. This is an advantage that our randomization approach offers.

4.3 Conditional Probabilities of Good Skills

Using an NSE threshold of 0.6, we calculated the conditional probability that a model having a certain skill score with respect to one variable (ET or streamflow) will be good at capturing the other variable (as described in Section 3.5). Figure 7a shows that the models that have good NSE_{ET} scores are likely to have good NSE_Q scores as well, indicated by a probability of 0.75 or more. Here, we can also see the discrepancies among the ET_{RS} products. None of the models were able to achieve $NSE_{ET} > 0.6$ against the MOD16A2 product. The highest NSE_{ET} range was 0.7, 0.8, and 0.85 for GLEAM, SSEBop and TerraClimate respectively. This result also reflects the varying agreement between the simulated ET from SWAT and different ET_{RS} products. Specifically, SWAT can generally capture ET_{RS} from TerraClimate better than others in our regions.

Conditional probabilities of having a good NSE_{ET} when NSE_Q is good are near zero for the GLEAM and MO16DA2 products (Figure 7b), for all ranges of NSE_Q. For the SSEBop and TerraClimate products, conditional probabilities are higher and generally increase with larger NSE_Q. However, the highest probabilities (when $NSE_Q \in (0.9, 0.95]$) are only around 0.65, much lower than those in Figure 7a. Thus, the probability that a model performing well for streamflow also does well for ET is quite low (relative to the probability of the converse case, that a model performing well for ET also does well for streamflow). This indicates that a model constrained by streamflow alone might not be able to reproduce a realistic ET estimate. Results from the conditional probability with KGE index show shows similar features but different in magnitudes with that of the NSE (Figure S3).

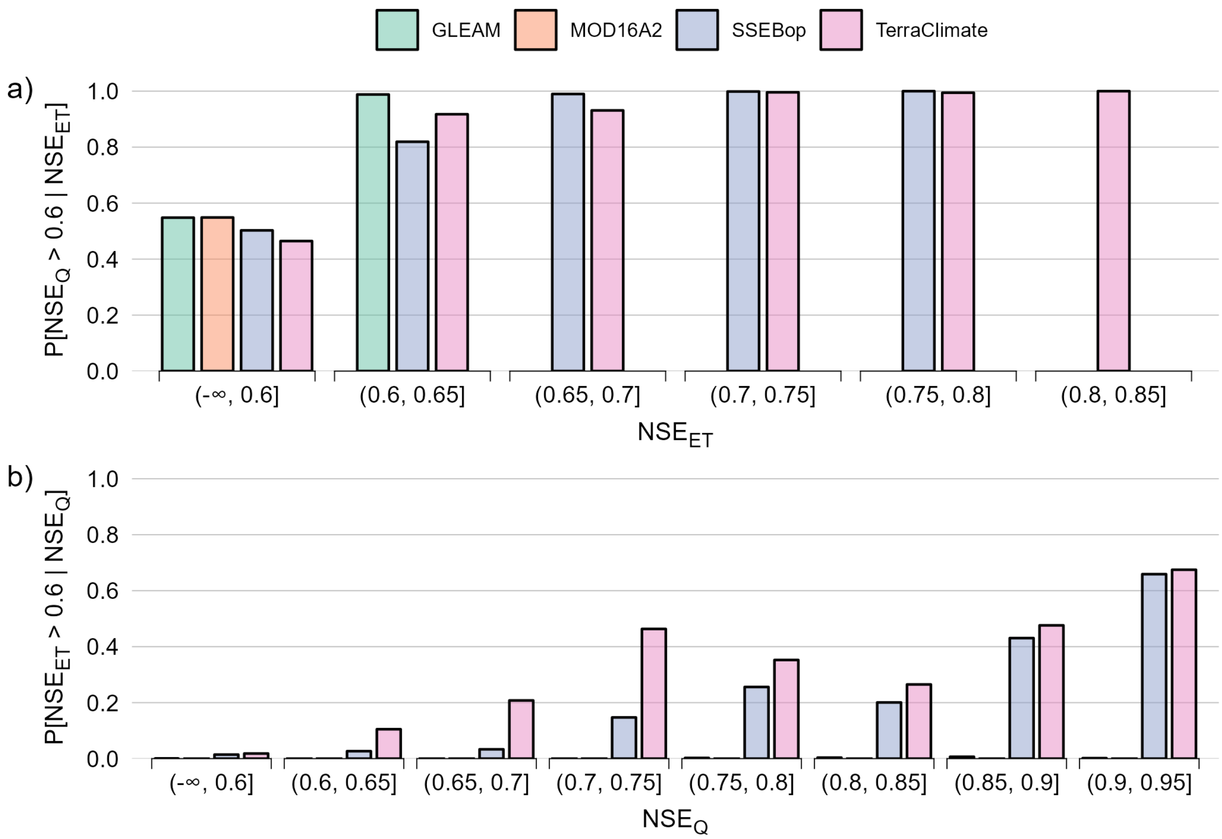

Figure 7. a) Conditional probability of having a good streamflow score ($NSE_Q > 0.6$) given a range of values of NSE_{ET}. b) Conditional probability of having a good ET score ($NSE_{ET} > 0.6$) given a range of values of NSE_Q. In panel a, some conditional probabilities, such as in the case of GLEAM when $NSE_{ET} > 0.7$, are not available because no models achieved the range of NSE_{ET} for the conditional probabilities to be calculated. In panel b, all probabilities are positive.

4.4 Implications for streamflow prediction using ET_{RS}, and limitations

Our findings suggest that prior to using ET_{RS} in model calibration, a randomized experiment, such as the one presented here, should be performed to explore the relationship between streamflow and ET skills. In areas where a negative correlation between model skills for ET and streamflow exists, the used of ET_{RS} products for streamflow estimation is in question especially considering the uncertainty in the accuracy of ET_{RS}. With the GLEAM and MOD16A2 products, we have demonstrated that the probabilities of having good model skills for streamflow is only observed within a certain range but not the best range model skill for ET. This means that trying to improve the model skill in simulating ET could lead to lower model skill for streamflow. The definition of behavioral model for streamflow prediction should corresponds to only a certain range but not the best range of model skill for ET. With all ET_{RS} products, we suggest using a behavioral range of model skill for streamflow estimation. Only using the best or a single good model skill for ET could results in a very uncertain model skill for streamflow, as the probability of having good model skill for streamflow when model skill for ET is good is not always 100%. This is in line with the concept of the equifinality thesis (Beven, 2006).

In ungauged catchments, the relationship between ET- and streamflow-skill is unknown. However, this might be inferred from neighboring gauged catchments with similar catchment characteristics. In addition, using a large sample of catchments for such a study could help to inferred the spatial pattern of the relation between model skill for ET and streamflow as well as the effect of catchment and meteorological characteristics on this relation. Furthermore, the approach proposed in this study can be combined with other parameter regionalization techniques (Hrachowitz et al., 2013; Razavi & Coulibaly, 2013), allowing a robust estimation of streamflow in ungauged catchments.

It is important to highlight a caveat in our investigation: ET_{RS} products used in this study are not “ground-truth”; rather, they were obtained from satellite images via algorithms and models with certain assumptions and limitations. Therefore, a low ET skill score does not necessarily mean that the model is bad in simulating ET. It simply means that the simulated ET from the model and the calculated ET from satellite images disagree, and both can be inaccurate. In regions where ET_{RS} products have been validated and shown to have high accuracies, they still can be used to improve streamflow estimation with more confidence.

5 Conclusions

Using seven catchments with diverse characteristics, and a large number of model runs with randomized parameters, we found that model parameters can influence model performance for streamflow and ET in different ways, thus there is no guarantee that a model that captures well one variable in calibration can perform well with respect to another variable. With certain ET_{RS} products (GLEAM and MOD16A2), the relationship between model performance with respect to streamflow and ET are asymmetric: models that perform well with ET are likely to perform well with streamflow, but not *vice versa*. Our results suggest that there are potential values in using remote sensing ET products for model calibration, but there is also a lot of uncertainty. This shed some light on the conflicting findings of earlier studies: depending on where the calibrated models landed on the spectrum of model skills, one may find using ET helpful or not helpful. A large-scale study with different types of models and a larger number of catchments spanning over more climatic and landscape characteristics is needed to pinpoint how catchment characteristics affect these different behaviors and the spatial patterns of the relation between model performance for streamflow and ET.

Acknowledgments

Hung Nguyen is supported by the Lamont-Doherty Postdoctoral Fellowship. Hong Xuan Do and Tu Hoang Le are supported by a grant for basic research (No. CSCB21-MTTN-07) from Nong

Lam University - Ho Chi Minh City. We thank Rohini Kumar for insightful comments on the initial version of the manuscript.

Open Research

Instrumental rainfall and streamflow data cannot be made public due to government regulations. Other input data and remotely sensed ET used in the project are in the public domain and are cited in Section 2. The code for running the SWAT model in R is available at https://doi.org/10.5281/zenodo.6569761.

References

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Scientific Data*, *5*. https://doi.org/10.1038/sdata.2017.191

Allen, R. G. (1986). A Penman for All Seasons. *Journal of Irrigation and Drainage Engineering*, *112*(4). https://doi.org/10.1061/(asce)0733-9437(1986)112:4(348)

Allen, R. G., Jensen, M. E., Wright, J. L., & Burman, R. D. (1989). Operational Estimates of Reference Evapotranspiration. *Agronomy Journal*, *81*(4). https://doi.org/10.2134/agronj1989.00021962008100040019x

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., Santhi, C., Harmel, R. D., Van Griensven, A., Van Liew, M. W., Kannan, N., & Jha, M. K. (2012). SWAT: Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4), 1491–1508. http://dx.doi.org/10.13031/2013.42256

Arnold, J. G., Srinivasan, R., Muttiah, R. S., & Williams, J. R. (1998). Large area hydrologic modeling and assessment part I: Model development. *Journal of the American Water Resources Association*, *34*(1), 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Beck, H. E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K., Giraldo-Osorio, J. D., & Xuan Thinh, N. (2020). RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements. *Remote Sensing of Environment*, *239*. https://doi.org/10.1016/j.rse.2019.111606

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2). https://doi.org/10.1016/j.jhydrol.2005.07.007

Crétaux, J. F., Arsen, A., Calmant, S., Kouraev, A., Vuglinski, V., Bergé-Nguyen, M., Gennero, M. C., Nino, F., Abarca Del Rio, R., Cazenave, A., & Maisongrande, P. (2011). SOLS: A lake database to monitor in the Near Real Time water level and storage variations from remote sensing data. *Advances in Space Research*, *47*(9). https://doi.org/10.1016/j.asr.2011.01.004

Dembélé, M., Ceperley, N., Zwart, S. J., Salvadore, E., Mariethoz, G., & Schaefli, B. (2020). Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, *143*, 103667. https://doi.org/10.1016/j.advwatres.2020.103667

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018).

Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *22*(2). https://doi.org/10.5194/hess-22-1299-2018

Dile, Y. T., Ayana, E. K., Worqlul, A. W., Xie, H., Srinivasan, R., Lefore, N., You, L., & Clarke, N. (2020). Evaluating satellite-based evapotranspiration estimates for hydrological applications in data-scarce regions: A case in Ethiopia. *Science of the Total Environment*, *743*. https://doi.org/10.1016/j.scitotenv.2020.140702

Do, H. X., Le, M. H., Pham, H. T., Le, T. H., & Nguyen, B. Q. (2022). Identifying hydrologic reference stations to understand changes in water resources across Vietnam-a data-driven approach. *Vietnam Journal of Earth Sciences*, *44*(1). https://doi.org/10.15625/2615-9783/16980

ESA. (2017). *Land Cover CCI Product User Guide Version 2. Tech. Rep.* http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

Fischer, G., Nachtergaele, F. O., Prieler, S., van Velthuizen, H., Verelst, L., & Wiberg, D. (2008). Global Agro-ecological Zones Assessment for Agriculture (GAEZ v1.2). *International Institute for Applied Systems Analysis and Food, Laxenburg, Austria and Agriculture Organization of the United Nation*.

Gui, Z., Liu, P., Cheng, L., Guo, S., Wang, H., & Zhang, L. (2019). Improving Runoff Prediction Using Remotely Sensed Actual Evapotranspiration during Rainless Periods. *Journal of Hydrologic Engineering*, *24*(12). https://doi.org/10.1061/(asce)he.1943-5584.0001856

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. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hall, D. K., Riggs, G. A., & Salomonson, V. V. (1995). Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data. *Remote Sensing of Environment*, *54*(2). https://doi.org/10.1016/0034-4257(95)00137-P

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., & Bayr, K. J. (2002). MODIS snow-cover products. *Remote Sensing of Environment*, *83*(1–2). https://doi.org/10.1016/S0034-4257(02)00095-0

Hargreaves, G. H., & Samani, Z. A. (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99. https://doi.org/10.13031/2013.26773

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., Anderson, M. C., Sadeghi, A. M., Hain, C. R., & Sharifi, A. (2018). Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology*, *556*, 39–49. https://doi.org/10.1016/j.jhydrol.2017.11.009

Hornáček, M., Wagner, W., Sabel, D., Truong, H. L., Snoeij, P., Hahmann, T., Diedrich, E., & Doubková, M. (2012). Potential for high resolution systematic global surface soil moisture retrieval via change detection using sentinel-1. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(4). https://doi.org/10.1109/JSTARS.2012.2190136

Hsu, K. L., Gao, X., Sorooshian, S., & Gupta, H. V. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, *36*(9). https://doi.org/10.1175/1520-0450(1997)036<1176:PEFRSI>2.0.CO;2

Huang, Q., Qin, G., Zhang, Y., Tang, Q., Liu, C., Xia, J., Chiew, F. H. S., & Post, D. (2020). Using Remote Sensing Data-Based Hydrological Model Calibrations for Predicting Runoff in Ungauged or Poorly Gauged Catchments. *Water Resources Research*, *56*(8). https://doi.org/10.1029/2020WR028205

Immerzeel, W. W., & Droogers, P. (2008). Calibration of a distributed hydrological model based on satellite evapotranspiration. *Journal of Hydrology*, *349*(3–4), 411–424. https://doi.org/10.1016/j.jhydrol.2007.11.017

Jiang, L., Wu, H., Tao, J., Kimball, J. S., Alfieri, L., & Chen, X. (2020). Satellite-based evapotranspiration in hydrological model calibration. *Remote Sensing*, *12*(3), 428. https://doi.org/10.3390/rs12030428

Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., Burrows, R. M., DelVecchia, A. G., Fritz, K. M., Shanafield, M., Burgin, A. J., Zimmer, M. A., Datry, T., Dodds, W. K., Jones, C. N., Mims, M. C., Franklin, C., Hammond, J. C., Zipper, S., Ward, A. S., … Olden, J. D. (2022). Assessing placement bias of the global river gauge network. *Nature Sustainability*, *5*(7), 586–592. https://doi.org/10.1038/s41893-022-00873-0

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. *Journal of Hydrology*, *535*. https://doi.org/10.1016/j.jhydrol.2016.02.018

Le, M. H., Nguyen, B. Q., Pham, H. T., Patil, A., Do, H. X., Ramsankaran, R. A. A. J., Bolten, J. D., & Lakshmi, V. (2022). Assimilation of SMAP Products for Improving Streamflow Simulations over Tropical Climate Region—Is Spatial Information More Important than Temporal Information? *Remote Sensing*, *14*(7). https://doi.org/10.3390/rs14071607

Liu, X., Liu, F. M., Wang, X. X., Li, X. D., Fan, Y. Y., Cai, S. X., & Ao, T. Q. (2017). Combining rainfall data from rain gauges and TRMM in hydrological modelling of Laotian data-sparse basins. *Applied Water Science*, *7*(3). https://doi.org/10.1007/s13201-015-0330-y

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., & Verhoest, N. E. C. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*(5). https://doi.org/10.5194/gmd-10-1903-2017

Monteith, J. L. (1965). Evaporation and environment. In *Symposia of the Society for Experimental Biology* (Vol. 19).

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900.

Mu, Q., Zhao, M., & Running, S. W. (2013). MODIS Global Terrestrial Evapotranspiration (ET) Product (MOD16A2/A3). *Algorithm Theoretical Basis Document*, *Collection*.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Neitsch, S. ., Arnold, J. ., Kiniry, J. ., & Williams, J. . (2011). Soil & Water Assessment Tool Theoretical Documentation Version 2009. *Texas Water Resources Institute*.

https://swat.tamu.edu/media/99192/swat2009-theory.pdf (accessed 10 March 2022)

Nguyen, D. N., & Nguyen, T. H. (2004). Climate and Climate Resources in Vietnam (in Vietnamese). *Agricultural Publishing House, Hanoi, Vietnam*.

Nguyen, T. V., Uniyal, B., Tran, D. A., & Pham, T. B. T. (2022). On the Evaluation of Both Spatial and Temporal Performance of Distributed Hydrological Models Using Remote Sensing Products. *Remote Sensing*, *14*(9), 1959. https://doi.org/10.3390/rs14091959

Nguyen, V. T., Dietrich, J., & Uniyal, B. (2020). Modeling interbasin groundwater flow in karst areas: Model development, application, and calibration strategy. *Environmental Modelling and Software*, *124*, 104606. https://doi.org/10.1016/j.envsoft.2019.104606

Nguyen, T. V, Dietrich, J., Dang, T. D., Tran, D. A., Van Doan, B., Sarrazin, F. J., Abbaspour, K., & Srinivasan, R. (2022). An interactive graphical interface tool for parameter calibration, sensitivity analysis, uncertainty analysis, and visualization for the Soil and Water Assessment Tool. *Environmental Modelling & Software*, *156*, 105497. https://doi.org/https://doi.org/10.1016/j.envsoft.2022.105497

Odusanya, A. E., Mehdi, B., Schürz, C., Oke, A. O., Awokola, O. S., Awomeso, J. A., Adejuwon, J. O., & Schulz, K. (2019). Multi-site calibration and validation of SWAT with satellite-based evapotranspiration in a data-sparse catchment in southwestern Nigeria. *Hydrology and Earth System Sciences*, *23*(2), 1113–1144. https://doi.org/10.5194/hess-23-1113-2019

Pan, S., Tian, H., Dangal, S. R. S., Yang, Q., Yang, J., Lu, C., Tao, B., Ren, W., & Ouyang, Z. (2015). Responses of global terrestrial evapotranspiration to climate change and increasing atmospheric CO2 in the 21st century. *Earth's Future*, *3*(1). https://doi.org/10.1002/2014EF000263

Parajuli, P. B., Jayakody, P., & Ouyang, Y. (2018). Evaluation of Using Remote Sensing Evapotranspiration Data in SWAT. *Water Resources Management*, *32*(3). https://doi.org/10.1007/s11269-017-1850-z

Phan, V. T., Ngo-Duc, T., & Ho, T. M. H. (2009). Seasonal and interannual variations of surface climate elements over Vietnam. *Climate Research*, *40*(1). https://doi.org/10.3354/cr00824

Priestley, C. H. B., & Taylor, R. J. (1972). On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Monthly Weather Review*, *100*(2). https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2

R Core Team. (2021). R core team (2021). In *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www. R-project. org*.

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. *Journal of Hydrology*, *567*, 668–683. https://doi.org/10.1016/j.jhydrol.2018.10.024

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C. J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., & Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3). https://doi.org/10.1175/BAMS-85-3-381

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis*, *22*(3). https://doi.org/10.1111/0272-4332.00040

Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H., & Verdin, J. P. (2013). Operational Evapotranspiration Mapping Using Remote Sensing and Weather Datasets: A New Parameterization for the SSEB Approach. *Journal of the American Water Resources Association*, *49*(3), 577–591. https://doi.org/10.1111/jawr.12057

Sirisena, T. A. J. G., Maskey, S., & Ranasinghe, R. (2020). Hydrological model calibration with streamflow and remote sensing based evapotranspiration data in a data poor basin. *Remote Sensing*, *12*(22). https://doi.org/10.3390/rs12223768

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3). https://doi.org/10.1016/S0378-4754(00)00270-6

Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F., & Watkins, M. M. (2004). GRACE measurements of mass variability in the Earth system. *Science*, *305*(5683). https://doi.org/10.1126/science.1099192

Tobin, K. J., & Bennett, M. E. (2017). Constraining SWAT Calibration with Remotely Sensed Evapotranspiration Data. *Journal of the American Water Resources Association*, *53*(3). https://doi.org/10.1111/1752-1688.12516

Tran, H., Nguyen, P., Ombadi, M., Hsu, K. L., Sorooshian, S., & Qing, X. (2019). A cloud-free modis snow cover dataset for the contiguous United States from 2000 to 2017. *Scientific Data*, *6*. https://doi.org/10.1038/sdata.2018.300

Trenberth, K. E., Fasullo, J. T., & Kiehl, J. (2009). Earth's global energy budget. *Bulletin of the American Meteorological Society*, *90*(3). https://doi.org/10.1175/2008BAMS2634.1

Willem Vervoort, R., Miechels, S. F., van Ogtrop, F. F., & Guillaume, J. H. A. (2014). Remotely sensed evapotranspiration to calibrate a lumped conceptual model: Pitfalls and opportunities. *Journal of Hydrology*, *519*(PD). https://doi.org/10.1016/j.jhydrol.2014.10.034

Zhang, Y., Chiew, F. H. S., Liu, C., Tang, Q., Xia, J., Tian, J., Kong, D., & Li, C. (2020). Can Remotely Sensed Actual Evapotranspiration Facilitate Hydrological Prediction in Ungauged Regions Without Runoff Calibration? *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019WR026236

Zhang, Y., Chiew, F. H. S., Zhang, L., & Li, H. (2009). Use of remotely sensed actual evapotranspiration to improve rainfall-runoff modeling in Southeast Australia. *Journal of Hydrometeorology*, *10*(4). https://doi.org/10.1175/2009JHM1061.1

Water Resources Resarch

Supporting Information for

Uncertain Benefits of Using Remotely Sensed Evapotranspiration for Streamflow Estimation–Insights from a Randomized, Large-Sample Experiment

Tam V. Nguyen[1], Hung T.T. Nguyen[2,*], Vinh Ngoc Tran[3], Manh-Hung Le[4,5], Binh Quang Nguyen[6], Hung T. Pham[6], Tu Hoang Le[7], Doan Van Binh[8], Thanh Duc Dang[9], Hoang Tran[10], Hong Xuan Do[11,12*]

[1]Department of Hydrogeology, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany

[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA

[3]Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109, USA

[4]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

[5]Science Applications International Corporation, Greenbelt, MD 20771, USA

[6]The University of Danang - University of Science and Technology, Da Nang, Vietnam

[7]Research Center for Climate Change, Nong Lam University – Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

[8]Master Program in Water Technology, Reuse and Management, Vietnamese German University, Ben Cat, Binh Duong 820000, Vietnam

[9]Department of Civil and Environmental Engineering, University of South Florida, Tampa, FL 33620, USA

[10]Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA 99354, USA

[11]Faculty of Environment and Natural Resources, Nong Lam University – Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

[12]Center for Technology Business Incubation, Nong Lam University – Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

*Corresponding author: Hung T.T. Nguyen (hnguyen@ldeo.columbia.edu) and Hong X. Do (doxuanhong@hcmuaf.edu.vn)

Contents of this file

Table S1.

Figures S1 to S3.

Table 1. List of different studies using ETRS for streamflow estimation and their findings.

References	ET_{RS} products	Stuy area/Model/Calibration technique	Key findings
Kunnath-Poovakka et al. (2016)	• CMRSET (ET) • AMSR-E (Soil moisture)	• Study area: 11 catchments in eastern Australia • Model: Simplified Australian Water Resource Assessment – Landscape model (AWRA-L) • Calibration technique: Shuffled Complex Evolution Uncertainty Algorithm with 15 different objective functions by combining the Root Mean Square Error and the correlation coefficient for ET and soil moisture	• Streamflow prediction in catchments with low average runoff can be improved using reliable ET products.
Demirel et al. (2018)	• Actual evapotranspiration (AET) based on MODIS data under cloud-free conditions	• Study area: The Skjern river basin in Denmark. • Model: mesoscale Hydrologic Model • Calibration technique: 7 behavioral simulations from the Shuffled complex evolution approach with the model performance criteria is the KGE for streamflow and SPAEF for AET. Three calibration scenarios: streamflow only, AET only, and both streamflow and AET	• Comparable model performance for streamflow in the case of streamflow calibration only and in the case of using both streamflow and AET for model calibration. • Much poorer model performance for streamflow for the case of AET-calibration only compare to the case of streamflow-calibration only.
Parajuli et al. (2018)	• SEBAL	• Study area: Big Sunflower River Watershed in Northwestern, Mississippi. • Model: SWAT • Calibration technique: Best parameter from SUFI-2 approach, three calibration scenarios: streamflow only, ET only, and both streamflow and ET	The streamflow-only and ET-only modeling scenarios showed equally good model performances for streamflow, followed by the flow-ET calibration scenario.
Rajib et al. (2018)	• MOD16A2	• Study area: Pipestem Creek watershed in North Dakota, United States • Model: modified SWAT • Calibration technique: SUFI-2 approach, 4 calibration scenarios: (M1) streamflow only, (M2) streamflow with biophysical parameters, (M3) streamflow and ET (lumped approach), (M4) streamflow and ET (distributed approach). KGE was used as the performance index for both streamflow and ET	• Including biophysical parameters (calibration scenario M2) slightly improve the model performance for ET and streamflow compared to that of M1 • Model performance for ET and streamflow in case of calibration scenario M3 increases compared to that of M1 and M2 for the validation increases • Model performance for ET and streamflow during the calibration period in the case of calibration scenario M3 is comparable with

References	ET_{RS} products	Stuy area/Model/Calibration technique	Key findings
			that of calibration scenarios M1 and M2. Model performance (M3) is the best for the validation period (among 4 calibrations scenarios)
Gui et al. (2019)	• NDVI-based ET algorithm	• Study area: 208 watersheds in the U.S. • Model: Xinanjiang model • Calibration technique: Three calibration scenarios: (1) streamflow only, (2) both streamflow and using both streamflow and ETRS for the entire period of record, (3) same as (2) but using ET_{RS} only during rainless periods. The optimal parameter set was determined by combining different optimization approaches.	• Lower model performance (mean NSE across 208 watershed) for streamflow was observed in calibration scenarios 2 and 3 compared to that of scenario 1.
Dembele et al. (2020)	• Twelve different ET_{RS} products	• Study area: Volta River basin, West Africa • Model: mesoscale Hydrologic Model (mHM) • Calibration technique: (1) Streamflow only and (2) 48 calibration scenarios as a combination of four distinct multivariate calibration strategies (the basin-average, pixel-wise, spatial bias accounting, and spatial bias-insensitive) using streamflow and ET.	• Adding ET_{RS} into the calibration scheme slightly tradeoff model performance for streamflow to improve the performance of the terrestrial water storage, temporal dynamics of soil moisture and spatila patterns of soil moisture.
Jiang et al. (2020)	• MOD16 ET	• Study area: 28 basins in the U.S. • Model: VIC • Calibration technique: Shuffled Complex Evolution, two calibration scenarios: (1) streamflow only, (2) spatial distributed ET calibration	• ET calibration yields better or similar streamflow performance in 29% of the basins compared to that from streamflow-based calibration,
Zhang et al. (2020)	• PLM-ET	• Study area: 222 basins in Australia • Model: Xinanjang and SIMHYD • Calibration technique: Genetic algorithm, four calibration scenarios: (1) streamflow-only, (2) ET only, (3) and (4) both ET and streamflow but with different objective functions.	• Model performance for streamflow in case of including ET_{RS} in the calibration (scenarios 2-4) calibration only is not as good as calibration against Q, especially in drier regions
Sirisena et al. (2020)	GLEAM ET	• Study area: four basins in the Chindwin River basin, Myanmar • Model: SWAT model • Calibration technique: three calibration scenarios: streamflow only, (2) ET only, and (3) both streamflow and ET	• In the single variable calibration scenarios (1 and 2), model performance for the targeted variable increases but for the other variable decreases. • Calibration that targets both ET and streamflow, acceptable model

References	ET_{RS} products	Stuy area/Model/Calibration technique	Key findings
			performance was achieved with both variables
Willem Vervoort et al. (2014)	• MOD16A3	• Study area: four catchments in New South Wales, Australia • Model: IHACRES • Calibration technique: shuffled complex evolution, three calibration scenarios: (1) streamflow only, (2) ET only, and (3) both streamflow and ET, results were compared with the case of using parameter regionalization and using ET_{RS} as direct model input.	• Calibration with ET and streamflow does not improve streamflow skills. Calibration against only ET is the worst, even worse than the parameter regionalization approach.

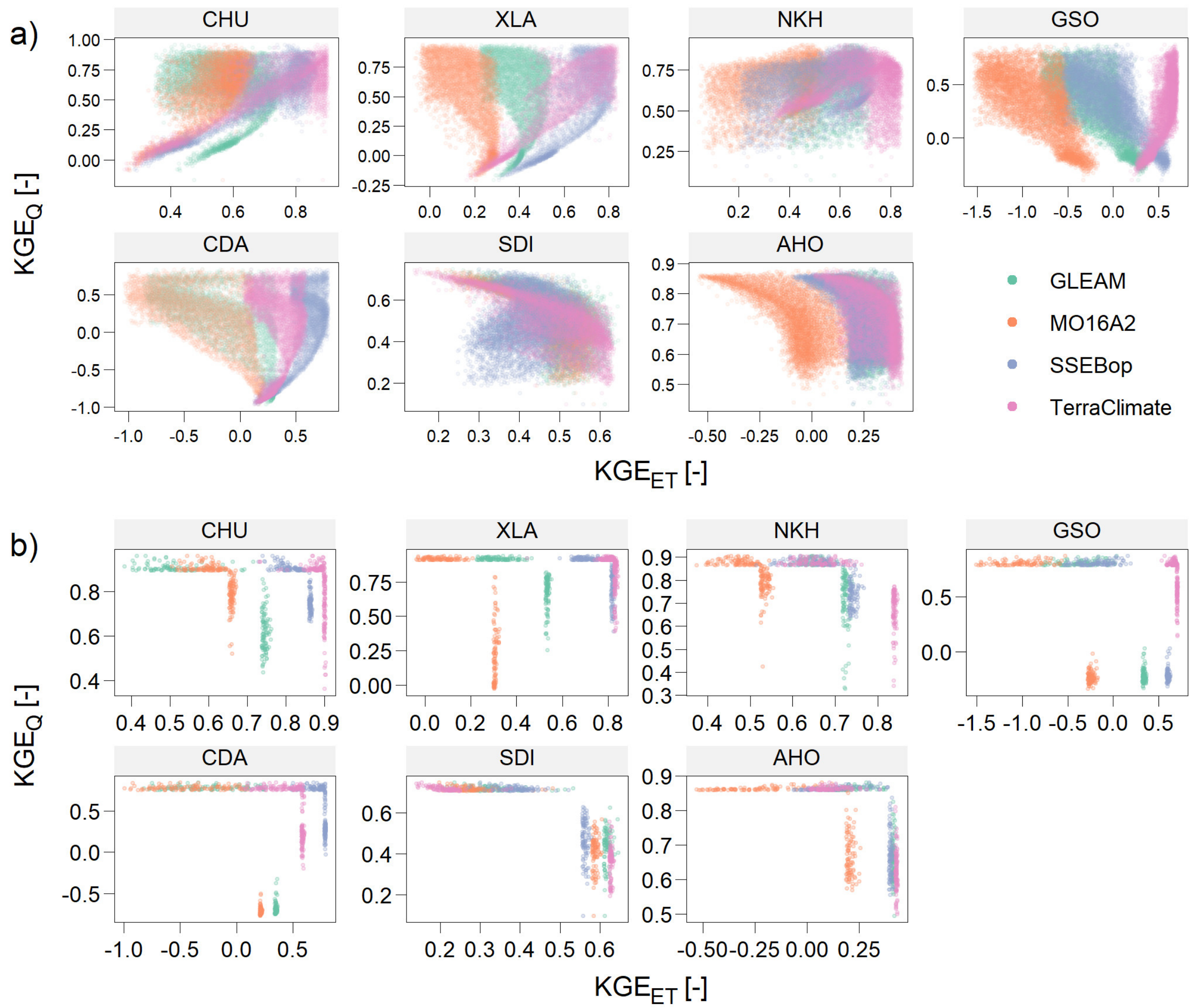

Figure S1. Distribution of KGE scores for ET (KGE_{ET}) versus KGE scores for streamflow (KGE_Q) for each catchment and ET_{RS} product. Panel a shows the scores of all 10,000 models and panel b shows the scores of models that are in either the top 100 for KGE_Q or the top 100 for KGE_{ET}. Note the large differences in x- and y-axis scales among the catchments.

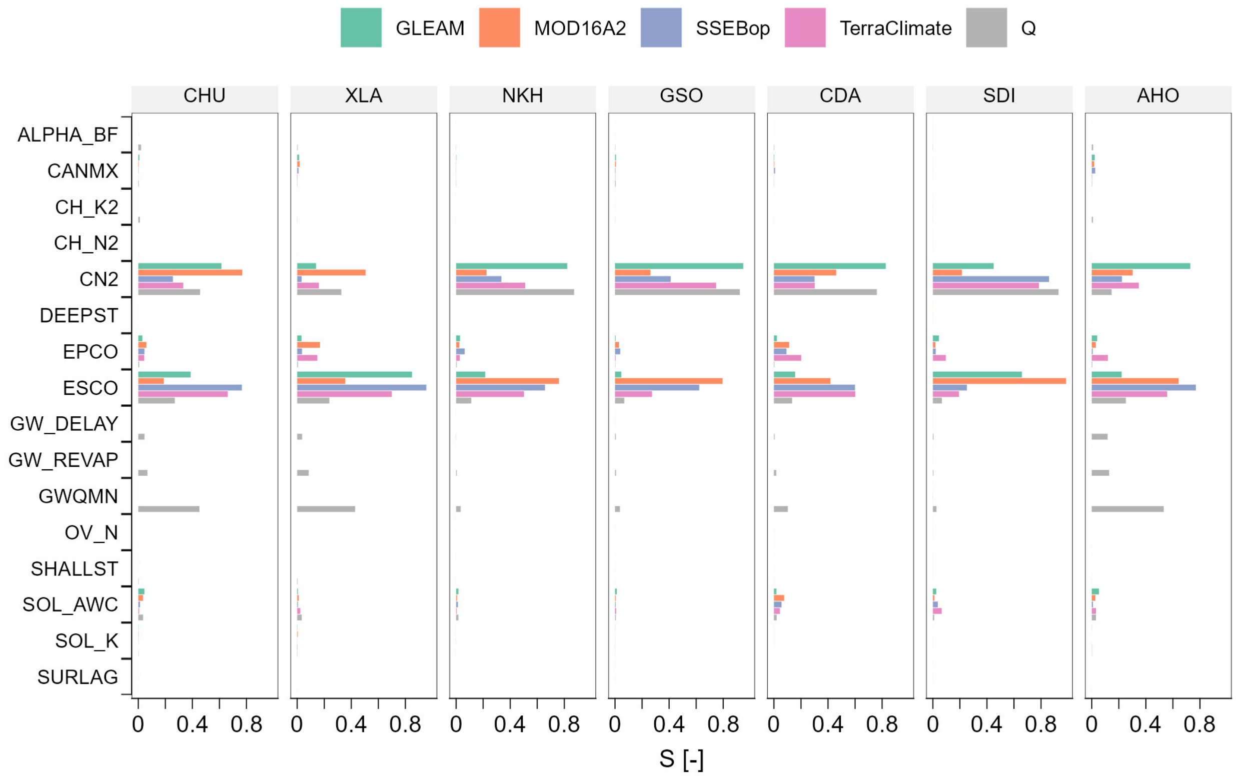

Figure S2. Total sensitivity (S) of streamflow and ET with respect to each model parameter in each catchment and variable (ET_{RS} product and observed streamflow Q). The objective function used in this analysis is the $KGE_{ET} + KGE_Q$.

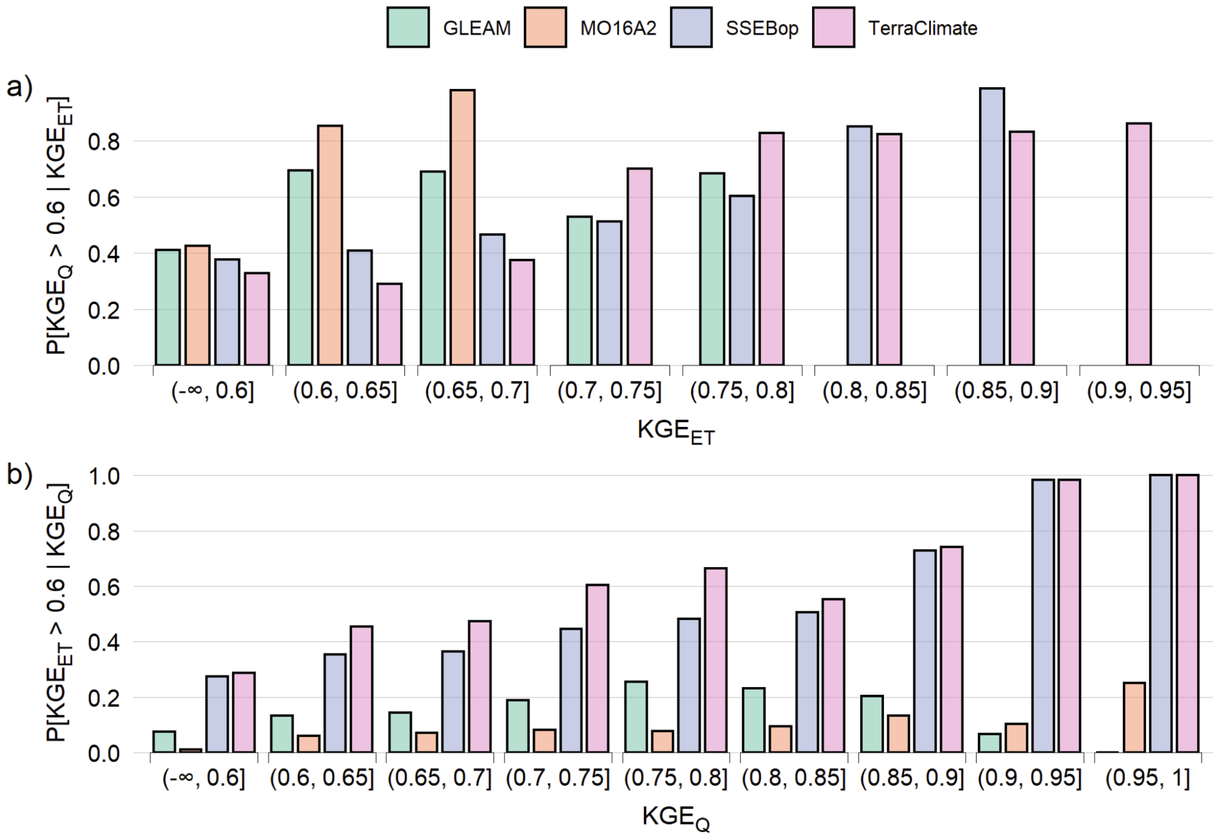

Figure S3. a) Conditional probability of having a good streamflow score ($KGE_Q > 0.6$) given a range of values of KGE_{ET}. b) Conditional probability of having a good ET score ($KGE_{ET} > 0.6$) given a range of values of KGE_Q.