Assessing the impact of PET estimation methods on hydrologic model performance

Dilhani Ishanka Jayathilake and Tyler Smith

ABSTRACT

Evapotranspiration is a necessary input and one of the most uncertain hydrologic variables for quantifying the water balance. Key to accurately predicting hydrologic processes, particularly under data scarcity, is the development of an understanding of the regional variation of the impact of potential evapotranspiration (PET) data inputs on model performance and parametrization. This study explores this impact using four different potential evapotranspiration products (of varying quality). For each data product, a lumped conceptual rainfall–runoff model (GR4J) is tested on a sample of 57 catchments included in the MOPEX data set. Monte Carlo sampling is performed, and the resulting parameter sets are analyzed to understand how the model responds to differences in the forcings. Test catchments are classified as energy- or water-limited using the Budyko framework and by eco-region, and the results are further analyzed. While model performance (and parameterization) in water-limited sites was found to be largely unaffected by the differences in the evapotranspiration inputs, in energy-limited sites model performance was impacted as model parameterizations were clearly sensitive to evapotranspiration inputs. The quality/reliability of PET data required to avoid negatively impacting rainfall–runoff model performance was found to vary primarily based on the water and energy availability of catchments.

Key words | Budyko classification, hydrologic model, potential evapotranspiration, uncertainty, variability

HIGHLIGHTS

- Model sensitivity to potential evapotranspiration (PET) errors was explored based on eco-regional and Budyko classifications.
- Although the model was not found to be sensitive to eco-region classification, the sensitivity varied along the water- to energy-limited continuum.
- This information, critically, can be used to better allocate limited resources for performing data collection and modeling and has benefits in data-scarce regions.

Dilhani Ishanka Jayathilake
Department of Physical & Environmental Sciences,
Texas A&M University-Corpus Christi,
Corpus Christi,
TX 78412,
USA

Tyler Smith (corresponding author)
Department of Civil and Environmental Engineering,
Clarkson University,
Potsdam,
NY 13699,
USA
E-mail: *tsmith@clarkson.edu*

doi: 10.2166/nh.2020.066

INTRODUCTION

Potential evapotranspiration (PET) is a key input to hydrological models. Yet, there is no consensus on which the PET estimation method is most applicable for hydrological modeling (Xu & Singh 2002; Oudin *et al.* 2005a). Verstraeten *et al.* (2008) presented a comprehensive overview of the scientific literature on methods for estimating PET and stated that the selection of one method from many is primarily dependent on the objectives of the study and the type of data available. However, if a perfect PET input was needed to run a hydrologic model successfully, such a model would be rarely used in the real world, as evapotranspiration is challenging/expensive to measure accurately. Therefore, understanding how models respond to PET inputs is critically important.

The influence of the applied PET method on rainfall–runoff models has been subjected to several studies. Parmele (1972) observed varying sensitivity to PET biases of +10% and −20% on three watershed models in nine US catchments, indicating that the relative importance of PET accuracy is likely to vary according to location, as well as how model performance is assessed (e.g., whether it focuses more on high or low flows). Vázquez (2003) used three different PET inputs to run the Système Hydrologique Européen (SHE) model and found that different PET inputs had significant impacts on the accuracy of streamflow simulation. Xu & Singh (2002) compared six PET estimation methods in Switzerland and found that radiation-based methods performed better than temperature-based methods. Evans (2003) tested several regional climate models associated with the CMD-IHACRES model on a catchment in the central USA and obtained optimal results when estimating evapotranspiration with temperature as a surrogate for PET. In contrast, Roderick *et al.* (2009a) discussed the downside of using a PET formula that does not adequately represent all relevant processes by highlighting recent trends in pan evaporation data. Over the past several decades, evaporation from Class-A pans has decreased, while at the same time, annual temperatures have risen (Fu *et al.* 2009; Roderick *et al.* 2009b). Therefore, they suggested that pan evaporation would have increased if rising temperature leads to rising evaporation; the opposite of what has been observed from recent pan data. Much of the observed declines in pan evaporation was attributed to declines in radiation and/or wind speed. This implies the potential importance of including other factors that are affecting evaporation, such as wind speed and radiation in the calculation of PET.

Although different PET methods may give significantly different results, some studies showed that hydrological models tend to be insensitive to differences in PET inputs (Andersson 1992; Andréassian *et al.* 2004; Oudin *et al.* 2005a). Andersson (1992) compared the output of the HBV model calibrated to seven different PET methods, including daily Penman PET values, mean monthly Penman PET values, and temperature-based PET models. The HBV model coped with the biases of PET methods by adapting its precipitation correction factor (i.e., a precipitation adjustment factor) without noticeable efficiency loss. While the differences between PET methods in terms of model efficiency were marginal, temperature-based methods slightly improved the performance of the model. According to Andersson (1992), mean Penman PET produced better runoff simulations than time-varying Penman PET. Nandakumar & Mein (1997) found that random PET errors could lead to significant runoff errors, but the sensitivity of watershed models depended on the dominant hydrological processes. While the bias in PET showed a linear relationship with bias in the predicted runoff, the bias in PET had a lower effect than bias in rainfall. Andréassian *et al.* (2004) tested the impact of a regionalized Penman PET on the performance of two rainfall–runoff models on a sample of 62 mountainous catchments. They found that the same average PET input for all catchments (i.e., a very simple assumption on PET) yielded the same result as a more complex input obtained from regionalization. Oudin *et al.* (2005b) compared different PET estimation methods on four conceptual rainfall–runoff models and found that both simplistic and complex PET estimation methods can achieve similar model performances. Oudin *et al.* (2005a) found no systematic improvements in rainfall–runoff model efficiencies when using temporally varying PET instead of mean PET and concluded that the rainfall–runoff models were poorly responsive to detailed PET inputs.

However, none of these studies has addressed rainfall–runoff model sensitivity to PET inputs based on catchment classification of long-term energy demands of watersheds (e.g., Budyko classification). We hypothesize that catchment energy availability represents a primary control on hydrologic model sensitivity to PET inputs. This study explicitly explores the importance of PET time series to hydrologic models in 57 catchments across diverse conditions, including climate, land cover, and energy/water-limitations due to differences in the dominant factors that influence the changes in the actual evapotranspiration (AET).

STUDY AREA AND DATA

A random subset of 57 catchments from the Model Parameter Estimation Experiment (MOPEX) data set (Duan *et al.* 2006) has been selected for analysis to provide broad geographical coverage with reasonable computational time. Test catchments are located within eight aggregated US EPA Level II eco-regions (Omernik 1987) in the USA (Figure 1) and have different climatic conditions (based on Köppen classification) ranging from humid subtropical to cold semi-arid (Beck *et al.* 2018). Study catchments range in the area from 80 to 10,000 km^2 (average ≈ 3,000 km^2), mean elevation from 10 to 2,600 m (average ≈ 590 m), and mean slope from 0.1 to 50% (average ≈ 9.7%). A number of recent studies have been based on MOPEX catchments (e.g., Sivapalan *et al.* 2011; Wang & Hejazi 2011; Carmona *et al.* 2014). This data set contains daily time series of precipitation, streamflow, maximum temperature, and minimum temperature, as well as a climatologic PET estimate that is based on pan evaporation data. Some of the catchments

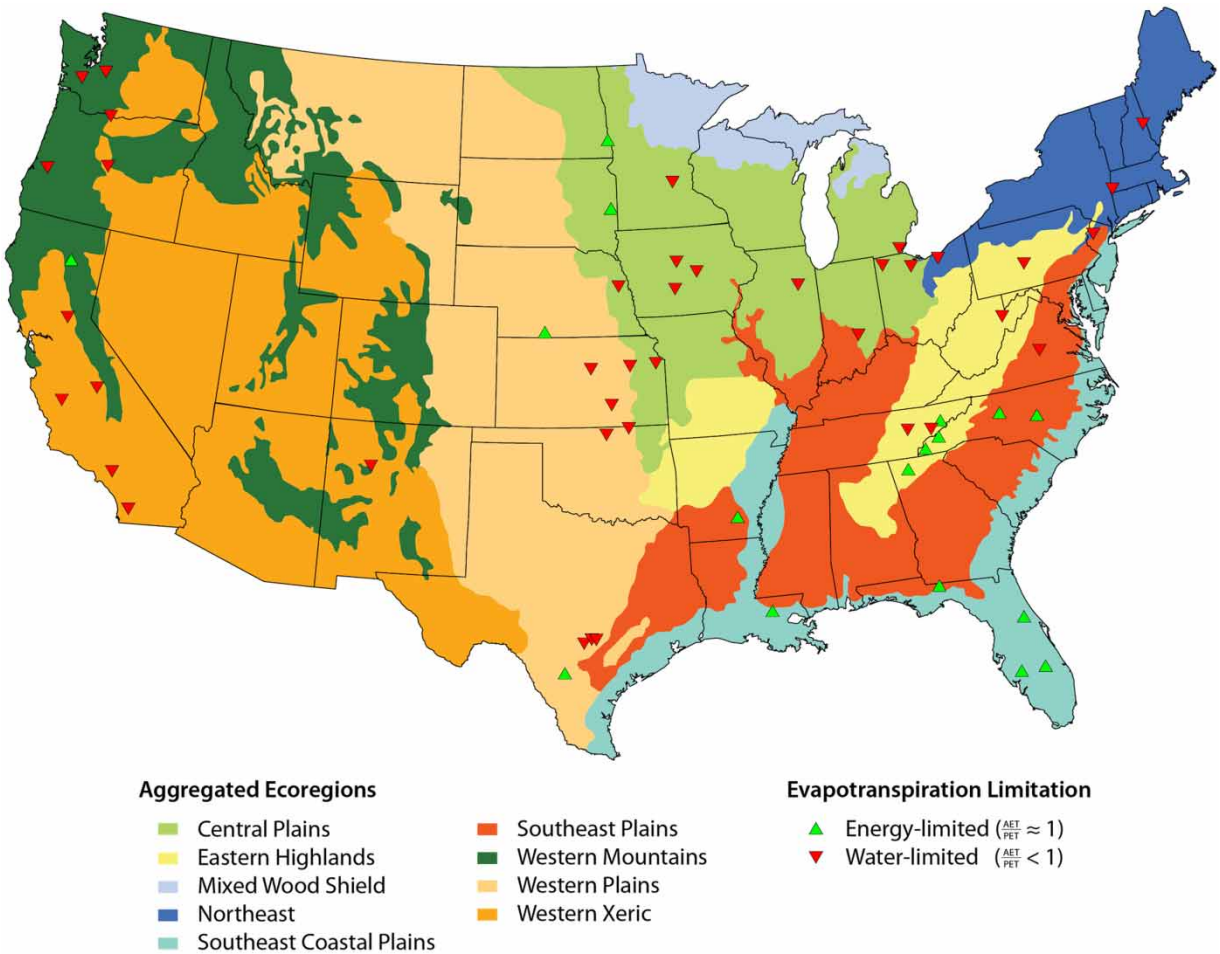

Figure 1 | A subset of 57 catchments from the MOPEX dataset located across United States (adapted from (Falcone 2011)).

included in the MOPEX data set contain periods of missing streamflow data. Therefore, the most recent 7 years of data with no missing streamflow days were selected for this analysis.

Four PET products were used in this study and are described below: (1) daily MOPEX, (2) mean monthly MOPEX, (3) Hargreaves and Samani (Hargreaves & Samani 1985), and (4) Hamon PET (Allen *et al.* 1998; Lu *et al.* 2005) (Figure 2). The daily PET data included in the MOPEX data set used for this analysis were calculated based on the National Oceanic and Atmospheric Administration (NOAA) Evaporation Atlas (Farnsworth *et al.* 1982). Daily PET data were produced by fitting monthly pan data taken from the NOAA Evaporation Atlas to a Fourier series with a single annual cycle (repeating each year). Daily values of mean PET data included in the MOPEX data set were averaged by month to obtain mean monthly PET data.

Temperature-based methods (i.e., Hargreaves-Samani (H-S) and Hamon methods) were chosen as temperature readings are often the most readily available meteorological data that relates to PET. Although the H-S method only uses a daily measurement of maximum and minimum temperatures as inputs, it effectively incorporates radiation indirectly. Relative humidity and cloudiness are not explicitly contained in the equation, but the difference in maximum and minimum air temperature is related to relative humidity and cloudiness (Samani & Pessarakli 1986). The H-S method has proven to produce good results and has great resilience in diverse climates around the world (Shahidian *et al.* 2012). Hamon's method (Allen *et al.* 1998; Lu *et al.* 2005) is a simple, empirical method that uses temperature as the major driving force for evapotranspiration, but also includes other variables such as daytime length and saturated vapor pressure. In this method, the daytime length is used as an index of the maximum possible incoming radiant energy and the saturated vapor pressure is the moisture-holding capacity of the air at the prevailing air temperature. It has been observed that Hamon's method produces similar estimates of PET as other more sophisticated methods (Vörösmarty *et al.* 1998).

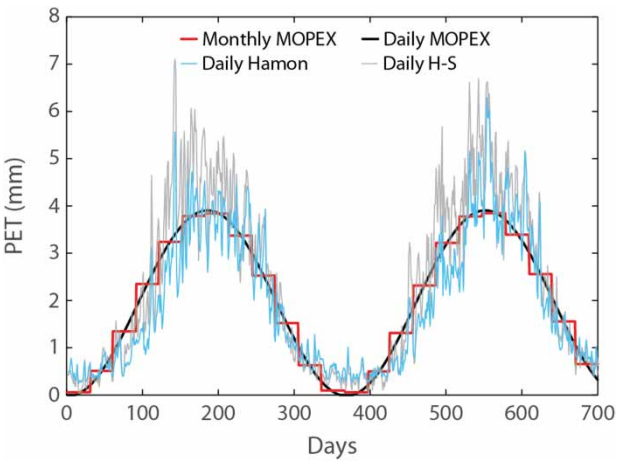

Figure 2 | Variability of PET estimated by Hargreaves-Samani and Hamon's methods, and daily and mean monthly pan PET for a representative catchment.

METHODS

Rainfall–runoff model

GR4J (Perrin *et al.* 2003), one of the most commonly applied, simple, conceptual rainfall–runoff models (e.g., Vaze *et al.* 2011; Coron *et al.* 2012; Smith *et al.* 2018), was selected for use in this study. The four-parameter GR4J model (Table 1) has been demonstrated to perform well in comparative studies (Perrin *et al.* 2001) and has been extensively tested on the range of climate conditions in the USA, Australia, France, etc. (Le Moine *et al.* 2007; Coron *et al.* 2012). The GR4J model involves two components for runoff simulation, a production component that regulates the water balance and a routing component that regulates the water transfer. The soil storage, characterized by its maximum capacity (X_1), stores rainfall and loses water as evapotranspiration and percolation. The production component classifies each day as either wet ($P > \text{PET}$; producing net rainfall) or dry ($P < \text{PET}$; producing a net evapotranspiration). Evapotranspiration from the production store relies on the wet/dry status of a day. For wet

Table 1 | Parameters of the GR4J model

Parameter	Description	Range
X_1	Capacity of the production store (mm)	[0, 1000]
X_2	Groundwater exchange coefficient (mm)	[−5, 5]
X_3	Capacity of the non-linear routing store (mm)	[0, 300]
X_4	Unit hydrograph time base (day)	[0.5, 5]

days, AET is set to be equal to PET, whereas for dry days, AET is modeled as the sum of the precipitation and the evapotranspiration from the store, which is modeled as a fraction of the net evapotranspiration depending on the water level in the production store relative to X_1. If there is net rainfall, a part of the rainfall will end in the production store. Percolation from the production store, together with the remaining net precipitation, reaches the routing component of the model. The total water available for routing is divided into two fractions: 90% comprise slow runoff that is routed through a unit hydrograph (UH_1) and a routing store, 10% are attributed to fast runoff that is routed through a unit hydrograph (UH_2). UH_1 and UH_2 utilize the same reference time base (X_4), which characterizes the rising time of the unit hydrographs. The amount of water that is stored in soil pores is defined by routing storage (X_3). The value of the routing store varies with the humidity and type of soil. The groundwater exchange coefficient (X_2) is a function of groundwater exchange that influences the routing store. While water inflow to the store from groundwater gives a positive value for X_2, water leaving from the store to groundwater gives a negative value.

The GR4J model was used in combination with the degree-day snowmelt module to account for snow in northern latitudes (Table 2; Kollat *et al.* 2012). The degree-day method assumes that precipitation occurs as snow when the air temperature drops below a threshold value (*TT*) and as rain otherwise. If precipitation occurs as snow, it is added to the dry snow component within the snowpack unless it ends up in the free water reservoir, which represents the liquid water content of the snowpack. The snowpack is capable of retaining melt water, expressed as a proportion of the total snowpack storage capacity (CWH). Thus, not all liquid water immediately results in runoff, rather the liquid water holding capacity of snow must first be exceeded. The melt water within the snow can also refreeze (if temperatures are below threshold *TT*) according to the refreezing parameter (*CFR*), which is expressed as a fraction of the degree-day factor (*CFMAX*).

Table 2 | Parameters of the degree-day snowmelt module

Parameter	Description	Range
TT	Threshold temperature (°C)	[−3, 3]
CFMAX	Degree-day factor (mm/°C)	[0, 20]
CFR	Refreezing factor (–)	[0, 1]
CWH	Water holding capacity of snowpack (–)	[0, 0.8]

Model calibration

Due to the conceptual nature of the GR4J model, calibration is required to determine appropriate parameter values. Candidate parameter sets were sampled under a Monte Carlo (MC) sampling framework (from uniform distributions) to generate 50,000 parameter sets, and the top 10% of parameter sets (i.e., 'behavioral' parameter sets), based on hydrograph fit, were retained for model parameter/performance analysis.

The GR4J parameters were calibrated to the Kling–Gupta Efficiency (KGE) objective function, which represents a weighting of three components that correspond to bias, correlation, and flow variability (Gupta *et al.* 2009). These components ensure that KGE is sensitive to errors in the overall distribution of streamflow. The formula for KGE is given as follows:

$$\mathrm{KGE} = 1 - \sqrt{(1-r)^2 + (1-\beta)^2 + (1-\Upsilon)^2} \tag{1}$$

$$\Upsilon = \frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}} \tag{2}$$

$$\beta = \frac{\mu_{\mathrm{sim}}}{\mu_{\mathrm{obs}}} \tag{3}$$

where r is the correlation coefficient, β is the bias ratio, Υ is the variability ratio, μ is the mean, and σ is the standard deviation. KGE ranges from $[-\infty, 1]$, with values closer to 1 representing better model fits.

Validation testing

Split-sample validation testing (Klemes 1986) was used to test model skill beyond the calibration period. For this study, 7 years of streamflow data for each catchment were used, with the first 4 years held for calibration and the last 3 years for validation. The model was validated using both direct-validation (i.e., calibrated and validated using the same PET forcing data) and using cross-validation (i.e., calibrated and validated using each PET forcing data).

Budyko classification

The selected MOPEX catchments were classified using the Budyko framework (Budyko 1974), which has been widely used to study basin-scale water and energy balances (Gerrits *et al.* 2009; Carmona *et al.* 2014). This framework assumes that long-term average annual evapotranspiration from a catchment is determined by water availability (i.e., rainfall) and energy availability (i.e., the maximum possible evapotranspiration). The theoretical Budyko curve (Figure 3) is given by two formulas that quantify the functional relationship between AET/P and PET/P:

$$\frac{AET}{PET} = \left\{1 + \left(\frac{P}{PET}\right)^{-\alpha}\right\}^{\frac{-1}{\alpha}} \tag{4}$$

$$\frac{Q}{PET} = \frac{P}{PET} - \left\{1 + \left(\frac{P}{PET}\right)^{-\alpha}\right\}^{\frac{-1}{\alpha}} \tag{5}$$

Test catchments were classified as energy-limited or water-limited based on their α value. The parameter α represents catchment similarity, while it is closely associated with the partitioning of water and energy availability of catchments (Gerrits *et al.* 2009; Carmona *et al.* 2014).

RESULTS

The GR4J model was tested with four different PET inputs: daily and mean monthly PET data from the MOPEX data set, daily H-S, and daily Hamon PET forcings. Budyko classification of the selected MOPEX catchments resulted in two hydrologically similar groups on the basis of magnitudes of the similarity parameter α (Carmona *et al.* 2014). In Figure 3, a catchment is energy-limited when the evaporative fraction (AET/PET) reaches 1 and AET approaches PET; a catchment is water-limited when AET/PET asymptotes to a value less than 1.

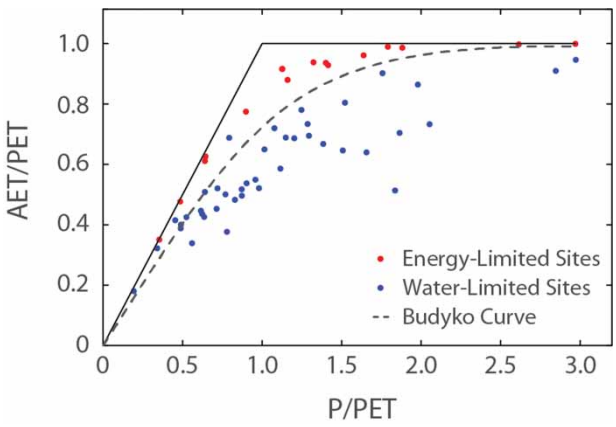

Figure 3 | Test catchments on Budyko curve based on α value.

The impact of PET inputs on the GR4J model under energy-limited and water-limited conditions was explored. The performance of the GR4J model in calibration and validation periods (in terms of KGE) were analyzed when directly validated with parameters obtained from the same PET forcing data set, as well as when cross-validated with parameters obtained from the other PET forcing data sets. Figure 4 shows the variation in the moving average of the median model performance of the top 10% of parameters (based on KGE) in cross-validation of each PET product for catchments fall along the water- to energy-limited continuum. The median model performance emphasizes the broader pattern along the water- to energy-limited continuum that is obscured by considering only the optimum parameter set. The spread between the lines (i.e., red and green) in Figure 4 is proportional to the impact of PET input data on model performance. Both daily MOPEX and mean monthly MOPEX PET data sets are based on the same actual pan PET data where mean monthly data is simply the monthly based averages of daily MOPEX data. Both H-S and Hamon methods are temperature-based methods where H-S is the most complex PET method among the two methods. Although the H-S method produces slightly higher PET volumes (Figure 2), in general, both PET methods exhibited similar patterns across test catchments (Figure 4). Thus, the model performance of the GR4J model for the most (i.e., H-S) and least (i.e., mean monthly) complex PET data sets are presented in detail (Figure 5). In Figure 5, the model performance for the top 10% of parameter sets is presented for three extremely water-limited (e.g., $\alpha = 1.26$) catchments and for three extremely energy-limited (e.g., $\alpha = 5$) catchments. The variability among the four boxplots (median, interquartile range) corresponds to the impact of different PET products on model performance. In addition to the model performance (in terms of KGE), modeled and observed hydrograph comparisons for representative catchments are presented for a

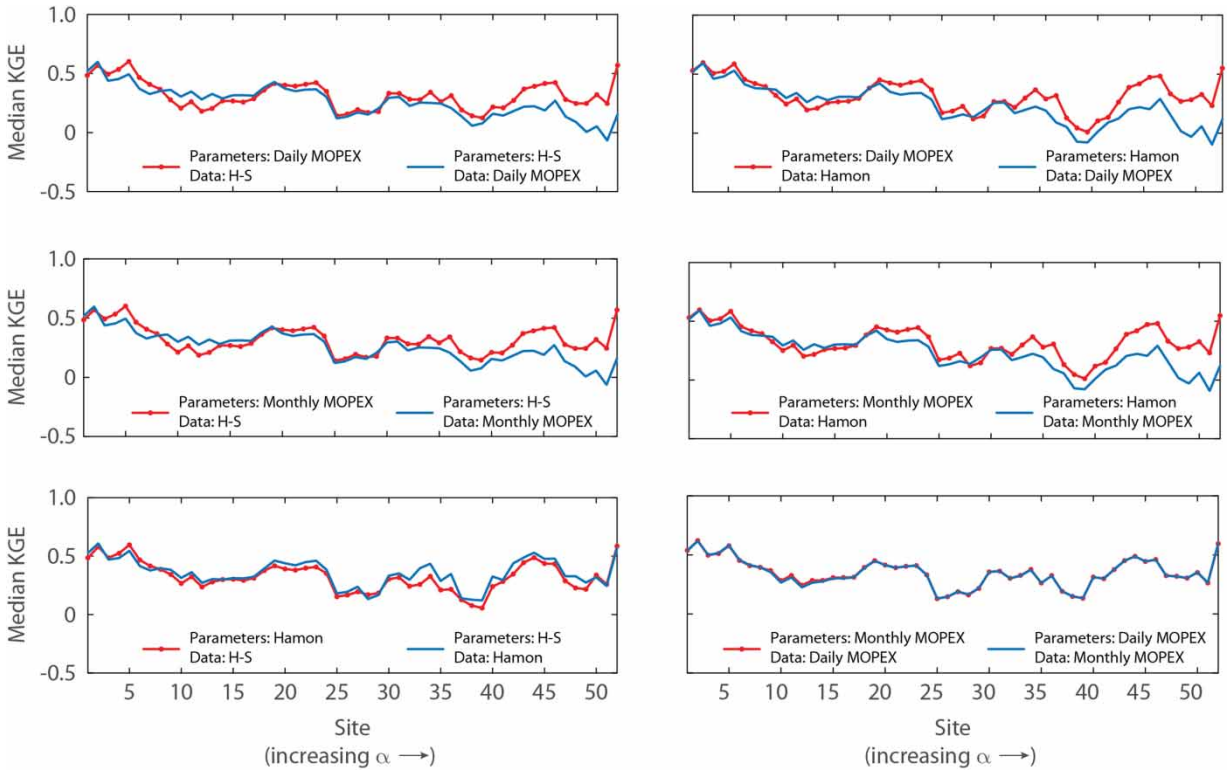

Figure 4 | Variation in median KGE values for cross-validation of each PET product in catchments sequenced in ascending order based on their α values (i.e., water- to energy-limited continuum). Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/nh.2020.066.

detailed comparison of model performance of the GR4J model (Figure 6).

The GR4J model was largely unaffected by the differences in PET forcings under water-limited conditions (Figures 5 and 6). It can be noted that model performance (as KGE) was nearly constant under both direct- and cross-validation tests. KGE was very similar regardless of PET input used, with median values and interquartile ranges showing little variation. Additionally, the spread between modeled and observed hydrographs for two cross-validated PET products showed small differences in water-limited catchments (Figure 6). In contrast, the model was impacted by the differences in PET inputs in most energy-limited catchments (Figures 4–6). The model was more strongly impacted by PET forcing inputs when cross-validated than direct-validated, as expected. KGE varied significantly (both in terms of median values and the interquartile ranges) between the four different PET inputs when cross-validated (Figures 4 and 5). The model performance varied significantly in many catchments under energy-limited conditions, when daily and monthly averaged MOPEX PET parameters were cross-validated to daily Hamon and H-S PET (Figures 4 and 5). Furthermore, model performance was lower (Figures 4 and 5) and the uncertainty (spread of boxplots in Figure 5) was larger when H-S or Hamon parameters were cross-validated on daily/mean MOPEX PET data sets than when daily/mean MOPEX parameters were cross-validated on H-S or Hamon PET data sets.

In many energy-limited catchments, the model performance exhibited less variability when cross-validated between H-S and Hamon PET data, as well as between daily MOPEX and monthly MOPEX PET (Figure 4). This can be attributed to the lack of significant difference between the PET products being cross-validated. Although we displayed model performance in detail for extreme cases (i.e., extremely water-limited and extremely energy-limited), the catchments classified in between the extreme endpoints exhibited intermediate sensitivity to PET reflective of their position along the energy- to water-limited continuum (Figure 4).

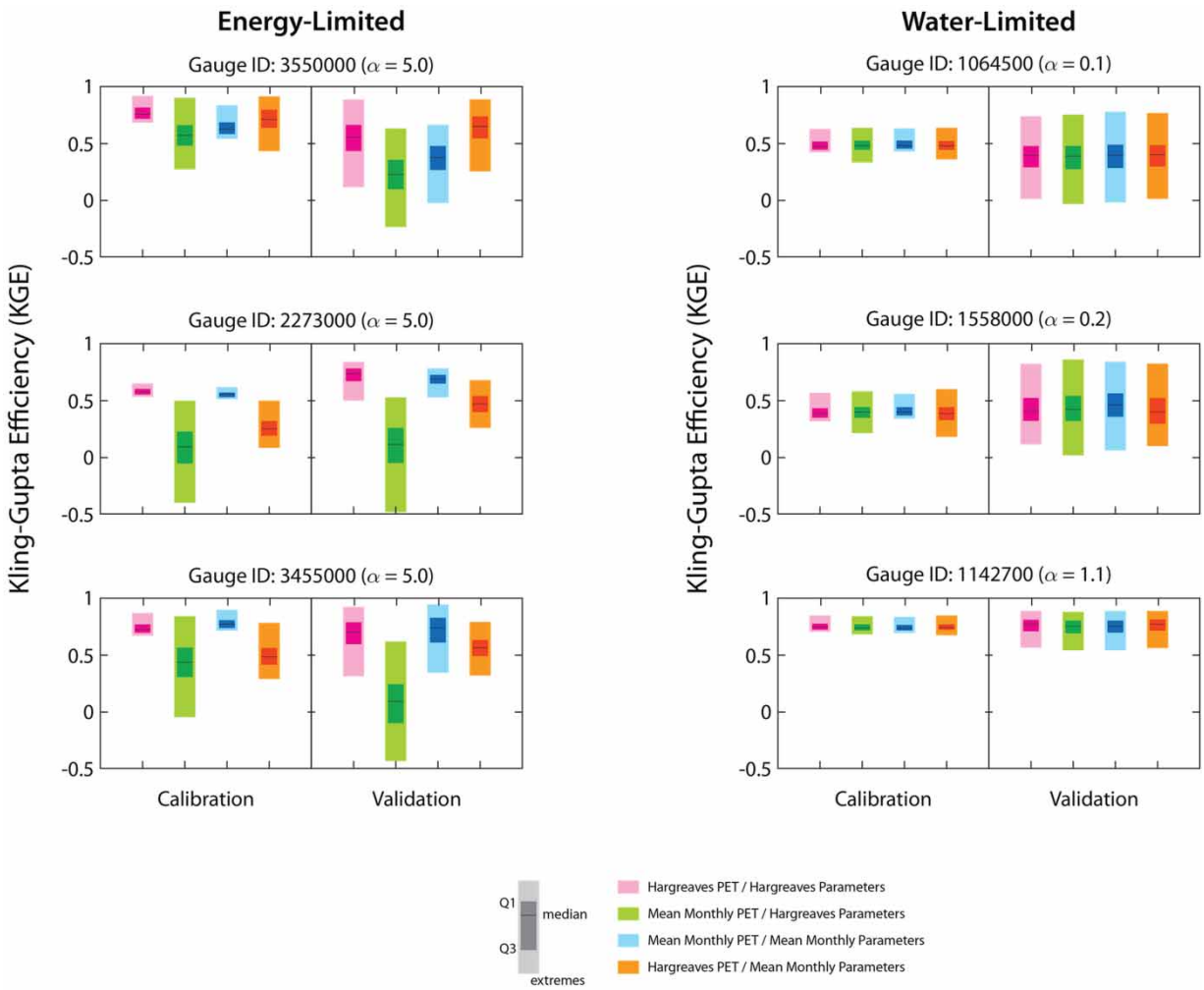

Figure 5 | Variation of model performance in water- and energy-limited catchments under direct- validation and cross-validation scenarios for different PET inputs.

The variability of GR4J model parameters to PET inputs was analyzed under energy-limited and water-limited conditions. Figure 7 presents the top 10% of parameters (based on KGE) of the model in detail for extreme cases (i.e., extremely water-limited and extremely energy-limited), when calibrated to each PET data set. Figure 8 exhibits the variation in medians of the most variable parameter along the water- to energy-limited continuum. The model parameterization of the GR4J model did not exhibit variability to changes in PET forcing data in water-limited catchments. For median values and behavioral parameter distributions of the GR4J parameters, the capacity of the production reservoir (X_1), the water exchange coefficient (X_2), nonlinear routing reservoir (X_3), and the unit hydrograph time base (X_4) were homogeneous across the four different PET inputs (Figure 7).

In contrast, the model parameterization was impacted by the differences in PET forcing data in most energy-limited catchments (Figures 7 and 8). The GR4J parameter describing the water exchange coefficient (X_2) exhibited the most variability. The median values and distribution of behavioral values of parameter X_2 varied greatly between PET inputs (e.g., energy-limited catchments shown in Figure 7). Other parameters of the GR4J model describing the capacity of the production reservoir (X_1), non-linear routing reservoir (X_3), and the unit hydrograph time base (X_4) did not exhibit significant variability.

The model parameters were not impacted by the differences in the temporal signal between daily and monthly

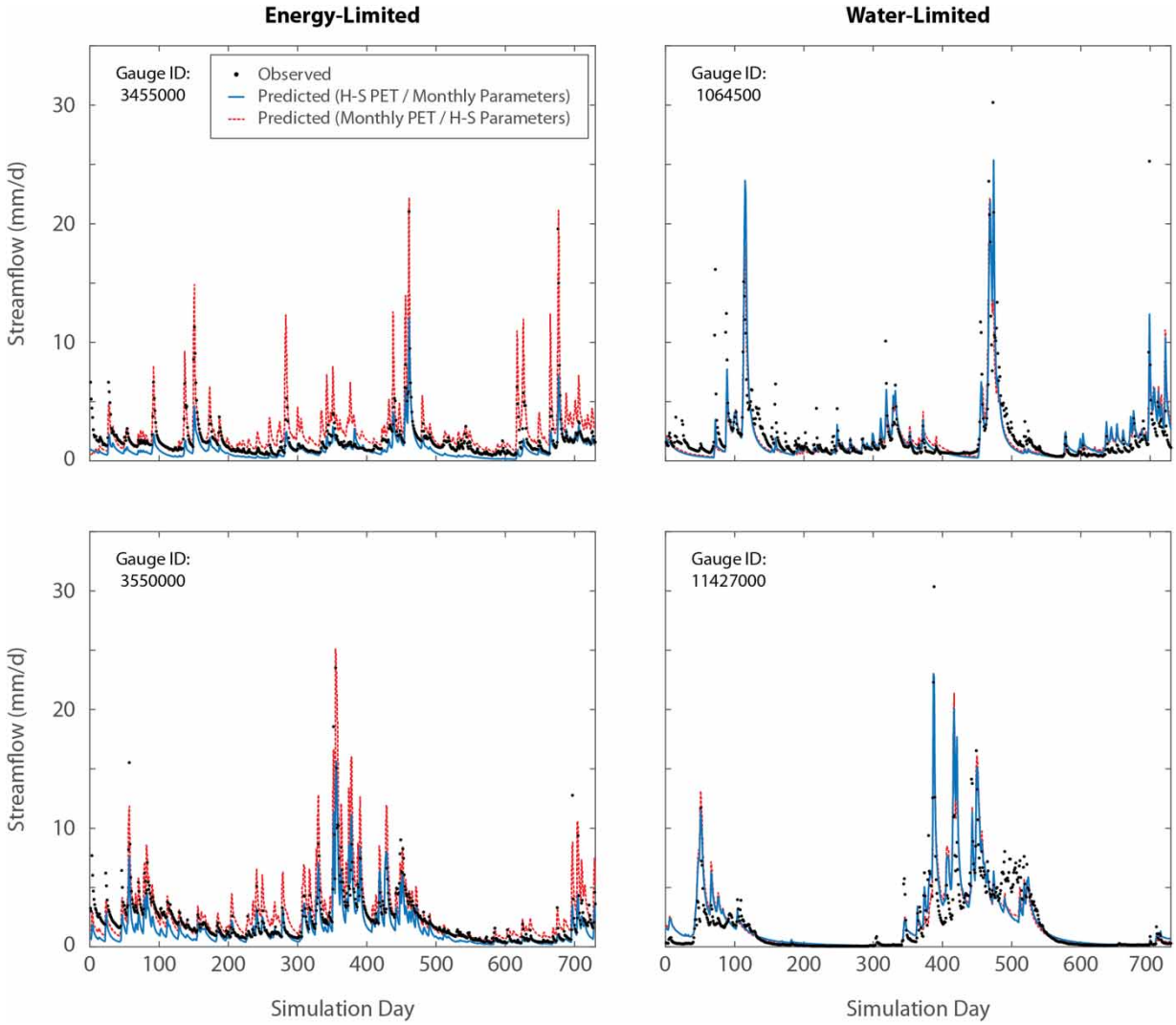

Figure 6 | Variation of modeled and observed streamflow time-series in water- and energy-limited catchments under cross-validation for different PET inputs.

mean MOPEX PET data sets. Medians of GR4J parameter X_2 in many energy-limited sites were similar (Figure 8). Moreover, model parameters were less impacted by the difference between H-S and Hamon PET data sets than the difference of those two data sets to daily and monthly mean MOPEX PET (Figure 8).

DISCUSSION

PET plays a major role in the long-term water balance in catchments. While the hydrological literature contains a wealth of PET models, little guidance exists on the importance of PET accuracy on streamflow simulations and how/if it varies across climatic regions. To better understand the relationship between PET inputs and streamflow modeling, the variability of the GR4J model performance and parametrization were tested on a sample of 57 catchments located in the coterminous US. We compared the parameterization and performance of the model across eco-regions and energy-limited/wet and water-limited/dry catchments.

Does the model exhibit variability by eco-region?

We tested 57 catchments in the USA covering eight contrasting ecological regions with distinct landform, climatic and

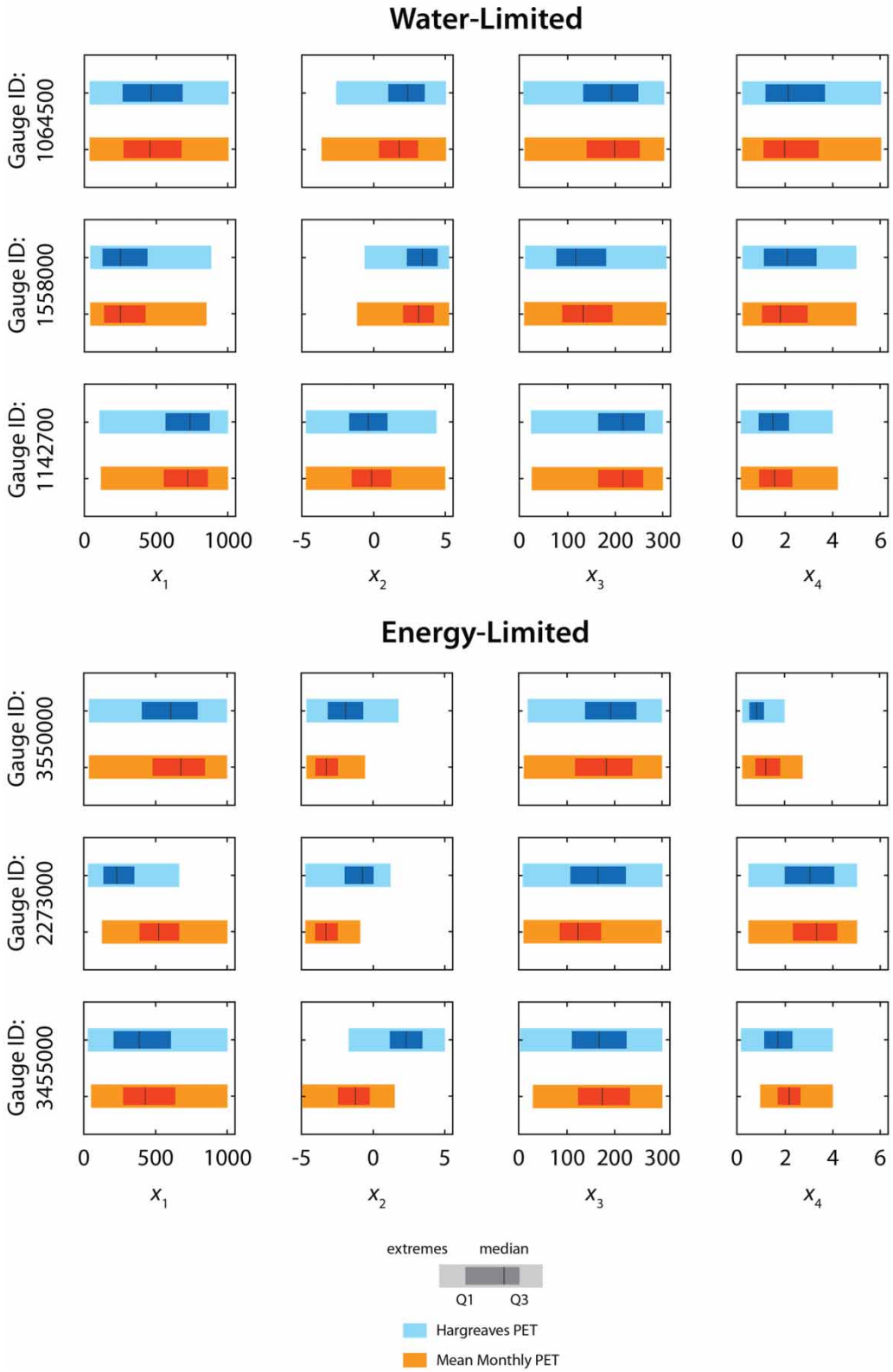

Figure 7 | Sensitivity of GR4J model parameters to different PET inputs in water-and energy-limited catchments.

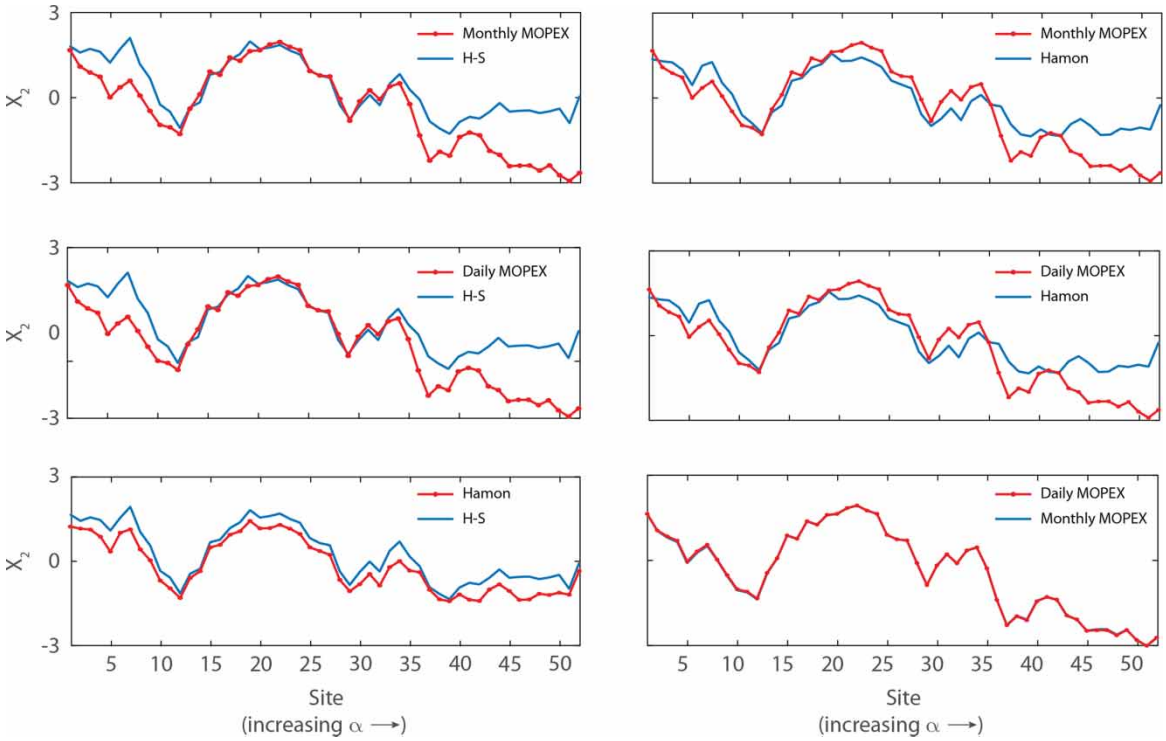

Figure 8 | Variation of GR4J parameter, water exchange coefficient (X_2) when calibrated to each PET input in catchments sequenced in ascending order based on their α values (i.e., water- to energy-limited continuum).

anthropogenic land use characteristics (Figure 1) in order to understand how the impact of PET inputs on the GR4J model varies geographically. Neither model performance nor model parameterization exhibited a strong pattern across eco-regions. The GR4J model uses a simplistic approach for converting PET to AET based on energy and soil water availability for evapotranspiration and is unable to explicitly account for variability in climate, vegetation, land use, and/or soil characteristics. As a result, the inability to identify relationships between model variability and catchment attributes across eco-regions is not surprising.

Does the model exhibit variability with energy or water availability?

The GR4J parameters and model performance were unaffected by the differences of PET inputs in water-limited catchments. Conversely, in energy-limited catchments, GR4J parameters and model performance exhibited sensitivity to PET forcing data sets when cross-validated with parameters obtained from the other PET forcing data set. In a water-limited catchment, the AET rate is less than the potential rate due to water-limitation (i.e., determined by the soil water availability). Therefore, a given change to the model forcing PET (e.g., daily MOPEX, mean monthly MOPEX, Hamon, and H-S) is unlikely to result in a significant change to the modeled AET and water-limited catchments are generally unaffected by PET inputs. In contrast, energy-limited catchments have no such water-limitation and as a result can often result in AET being equal to PET (at which point energy is limiting). When this occurs, the difference in PET inputs to an energy-limited catchment is likely to result in a potentially significant difference in the modeled AET. Thus, energy-limited catchments are expected to be largely impacted by PET forcings. Given the differences in the limiting factors that influence the changes in AET in the water- and energy-limited regions, model sensitivity to PET forcings differ between water- and energy-limited catchments. As such, the general findings of many earlier studies (Andréassian *et al.* 2004; Oudin *et al.* 2005a) that concluded hydrologic models to be insensitive to PET inputs can likely be

explained by locating their test catchments along the water- to energy-limited continuum.

How are the GR4J model parameters impacted by PET inputs?

The GR4J model parameters were impacted by PET forcing data in energy-limited catchments, where the groundwater exchange coefficient (X_2), a water balance parameter, showed the most significant change. Variability of the water balance parameter to changes in PET amount is reasonable, as it represents a control on the partitioning of water between stores in the model. When X_2 is negative, water is lost to deep groundwater; when positive, water is gained from groundwater to the basin. Therefore, the value of X_2 varies in response to changes in PET volume (to maintain the best fit to the hydrograph). However, the other water balance parameter of the GR4J model, the capacity of the production store (X_1), did not vary significantly with PET inputs. The routing parameters of the model, the capacity of the non-linear routing reservoir (X_3), and the unit hydrograph time base (X_4) were not responsive to differences in PET inputs. Equifinality, parameter sampling/sample size, and/or the threshold used to select parameter sets in the analysis can confound our understanding of the parameters.

How does the validation testing approach impact our understanding of model variability?

Under direct-validation, the predictive performance of the model is tested against the same characteristic PET data set that was used for the model calibration (i.e., assuming the PET data set is representative of the true condition). In contrast, under cross-validation, the performance of the model is tested against an independent PET data set that is not used for model calibration (i.e., assuming that the forcing PET is not necessarily representative of the true condition). Thus, while direct-validation tests the accuracy of the model, cross-validation tests the robustness of the model to variations in input data. As such, the model exhibits more variability with the PET differences when cross-validated with parameters obtained from the other PET forcing data than when direct-validated with parameters obtained from the same characteristic PET forcing data.

Most energy-limited catchments exhibited the highest model performance variability when cross-validated between daily/mean MOPEX PET and Hamon or H-S PET data sets due to increased variability in model parameters. However, cross-validation combinations of daily and mean MOPEX PET estimates and H-S and Hamon PET did not cause significant changes in the model performances or in the parameter values. Given the manner in which the daily/mean MOPEX PET data sets were derived (see section 'Study area and data'), it is clear that both are based on long-term average values with the same PET series repeated each year. Thus, model parameter values were homogeneous between the two data sets, resulting in similar model performance. Both H-S and Hamon models were based on daily temperature data which can result in similar PET patterns. This can be attributed to the limited difference in model parameters that resulted in similar model performances.

Does the directionality of cross-validation matter?

In many energy-limited catchments, the model exhibited greater performance degradation when H-S or Hamon parameters were cross-validated to daily/mean MOPEX PET data sets than when daily/mean MOPEX parameters were cross-validated to H-S or Hamon PET data sets. H-S and Hamon PET forcing data sets represent the daily temporal dynamics of PET, while the daily/mean MOPEX PET data sets represent long-term averages (Figure 2). When a model is calibrated to a specific, complex signal (e.g., fine temporal dynamics) and tested on a much more generic signal (i.e., long-term averages), the model is less likely to reproduce the pattern than when it is calibrated to the generic signal (which underlies the complex signal) and validated on the complex signal. Over-fitting to noise in a data signal can negatively impact on the performance by decreasing the models' ability to 'generalize'.

What is the impact of differences in the magnitude of PET inputs?

In this study, the daily and mean monthly MOPEX PET data sets had smaller long-term PET volumes than the H-S and Hamon PET data sets. To understand the impact of this on model parameterization and performance, the mean

monthly data set was synthetically shifted such that its long-term average volume was equal to the H-S data set (without altering the dynamics of its signal). The GR4J model was re-calibrated using the shifted mean monthly data set and model performance and parameterization of the original mean monthly, the shifted mean monthly, and the H-S PET data sets were analyzed (Figure 9). Although the model performance and parameterization varied between the original mean MOPEX and the H-S data sets, the shifted mean PET data set showed a similar model performance and parameterization to H-S PET under both direct- and cross-validation scenarios. The GR4J model was more strongly impacted by the magnitude of the long-term mean of the PET than from differences in the temporal signal. Thus, in hydrologic models, the accuracy of long-term average PET is more important than the specific PET signals with respect to the performance and parameterization. The effects of this may be obscured in water-limited sites as evapotranspiration there is constrained by water availability. However, the accuracy of long-term PET dynamics is particularly important over regions where evapotranspiration is limited by the availability of energy, where this study showed increased PET-driven hydrologic model variability.

CONCLUSIONS

The purpose of this study was to understand the sensitivity of hydrologic model predictions to PET forcing data. It was demonstrated, across 57 test catchments (Figure 1) and four different PET products of varying sophistication (Figure 2), that hydrologic model sensitivity to PET inputs is based on the energy availability of the catchment. As a result, the disparate conclusions of previous studies can be unified under a common framework of understanding; namely, that the relationship between PET and AET is based on water and energy availability and the sensitivity of hydrologic models to PET are a result of the propagation of this threshold mechanism through the model itself. Critically, this information can be used to better allocate limited resources for performing data collection and modeling and has particular benefits in data-scarce regions.

The impact of PET inputs on the GR4J model did not exhibit a spatial pattern based on eco-region classification; however, the modeling results were sensitive to the energy availability of catchments based on a Budyko classification (Figure 3). In energy-limited catchments, the GR4J model responded to changes in PET inputs through its parameters, primarily in terms of the groundwater exchange coefficient (X_2). Consequently, the performance of the GR4J model varied with PET forcing data sets. While rainfall–runoff models are likely to be most strongly influenced by PET forcing data in wet/energy-limited catchments, in dry/water-limited catchments significant variability was not observed.

The GR4J model exhibited the highest variability when cross-validating between daily or mean MOPEX PET and daily Hamon or H-S data sets. The accuracy of long-term PET magnitude was more influential on model parameters

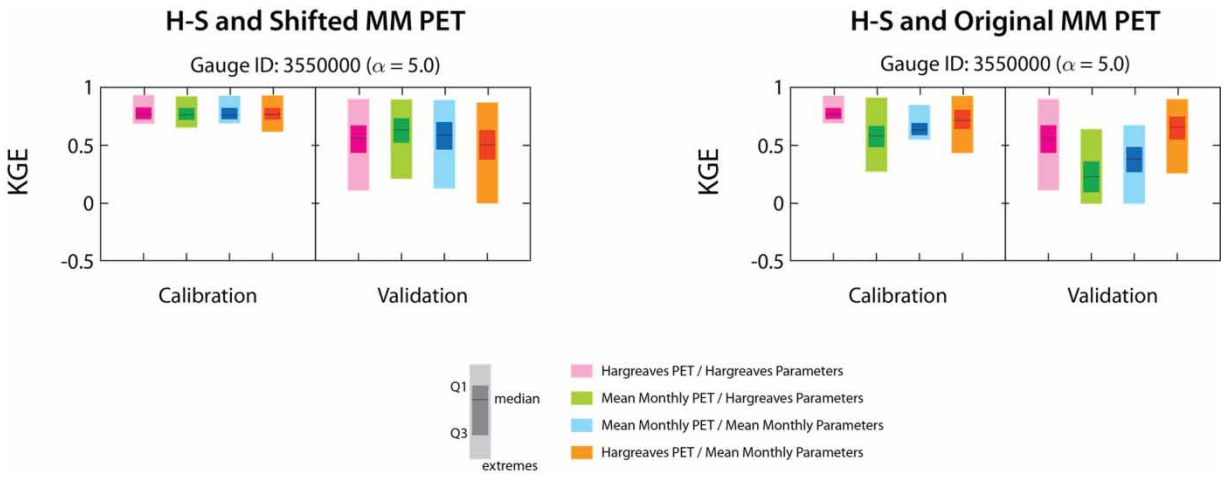

Figure 9 | Sensitivity of GR4J model performance to different PET inputs.

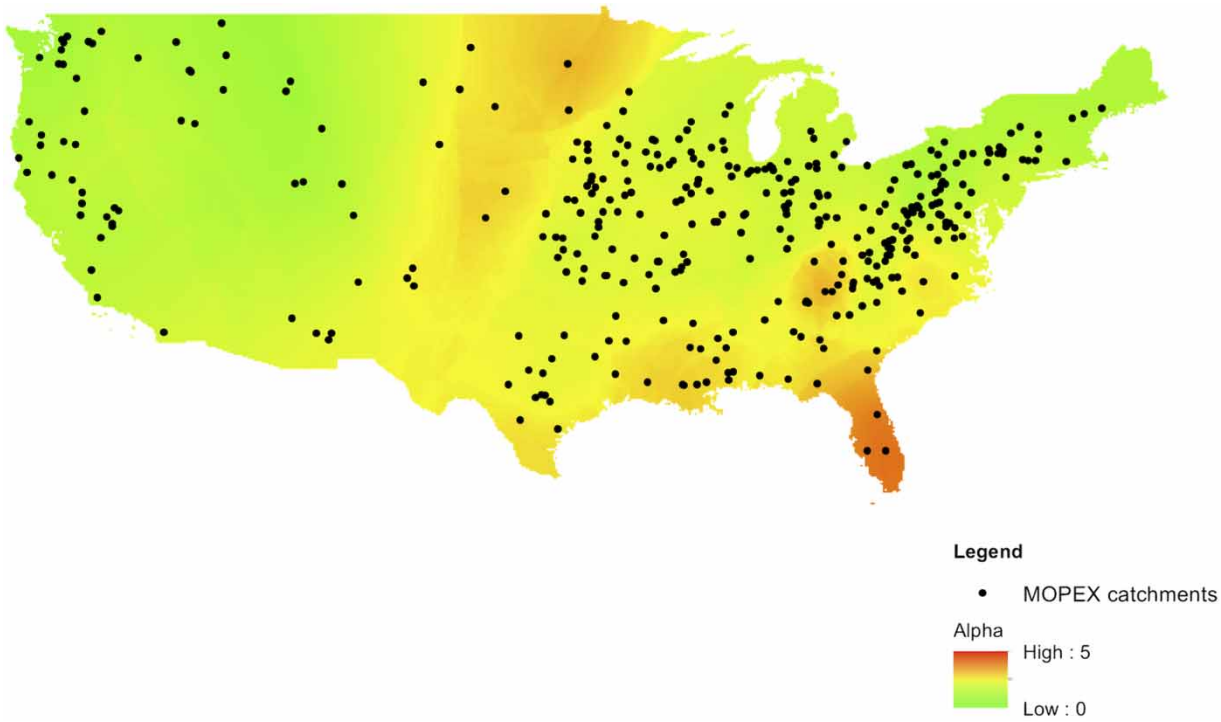

Figure 10 | A heat map showing variation in energy- and water-limitation across United States based on alpha values for 438 MOPEX sites.

and ultimately on model performance than the accuracy of temporal signal. The quality of PET data should not be ignored when catchments are under energy-limited conditions. More realistic PET forcing data should be used to better constrain model parameters sensitive to the PET/AET process and improve model performance, transferability, and fidelity. The results from this study can serve as a guide for data collection and resource allocation efforts in support of hydrologic modeling based on where catchments fall along the water- to energy-limited continuum (Figure 10).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available through the National Oceanic and Atmospheric Administration's Model Parameter Estimation Experiment (MOPEX) at ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/.

ACKNOWLEDGEMENT

Funding for this research was provided through a graduate scholarship from Clarkson University awarded to D. I. J.

REFERENCES

Allen, R. G., Pereira, L. S., Raes, D., Smith, M. & FAO 1998 *Crop Evapotranspiration – Guidelines for Computing Crop Water Requirements – FAO Irrigation and Drainage Paper 56*. Irrigation and Drainage, pp. 1–15. https://doi.org/10.1016/j.eja.2010.12.001

Andersson, L. 1992 Improvements of runoff models - what way to go? *Nordic Hydrology* **23** (5), 315–315.

Andréassian, V., Perrin, C. & Michel, C. 2004 Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models. *Journal of Hydrology* **286** (1–4), 19–35. https://doi.org/10.1016/j.jhydrol.2003.09.030.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A. & Wood, E. F. 2018 Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data* **5**, 180214.

Budyko, M. I. 1974 Climate and life/by M. I. Budyko. English ed. edited by David H. Miller. - Version details – Trove. Available from: https://www.mendeley.com/research-papers/climate-life-m-i-budyko-english-ed-edited-david-h-miller-version-details-trove/?utm_source=desktop&utm_medium=1.17.13&utm_campaign=open_catalog&userDocumentId=%7B72f22722-b18a-3f42-9ad2-ffeba1067c93%7D (Retrieved 19 March 2018).

Carmona, A. M., Sivapalan, M., Yaeger, M. A. & Poveda, G. 2014 Regional patterns of interannual variability of catchment water balances across the continental U.S.: a Budyko framework. *Water Resources Research* **50** (12), 9177–9193. https://doi.org/10.1002/2014WR016013.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M. & Hendrickx, F. 2012 Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. *Water Resources Research* **48** (5). https://doi.org/10.1029/2011WR011721

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L. & Hogue, T. 2006 Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. *Journal of Hydrology* **320** (1–2), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031.

Evans, J. P. 2003 Improving the characteristics of streamflow modeled by regional climate models. *Journal of Hydrology* **284** (1–4), 211–227. https://doi.org/10.1016/j.jhydrol.2003.08.003.

Falcone, J. A. 2011 *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow*. Reston, VA. Available from: http://pubs.er.usgs.gov/publication/70046617

Farnsworth, R. K., Thompson, E. S., Baldrige, M. & Byrne, J. V. 1982 *Mean Monthly, Seasonal, and Annual Pan Evaporation for the United States National Oceanic and Atmospheric Administration National Oceanic and Atmospheric Administration Climate Database Modernization Program*. Available from: http://www.dynsystem.com/netstorm/docs/NWS34EvapTables.pdf.

Fu, G., Charles, S. P. & Yu, J. 2009 A critical overview of pan evaporation trends over the last 50 years. *Climatic Change* **97** (1–2), 193–214. https://doi.org/10.1007/s10584-009-9579-1.

Gerrits, A. M. J., Savenije, H. H. G., Veling, E. J. M. & Pfister, L. 2009 Analytical derivation of the Budyko curve based on rainfall characteristics and a simple evaporation model. *Water Resources Research* **45** (4). https://doi.org/10.1029/2008WR007308

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.

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.

Klemes, V. 1986 Operational testing of hydrological simulation models. *Hydrological Sciences Journal* **31** (1), 13–24. https://doi.org/10.1080/02626668609491024.

Kollat, J. B., Reed, P. M. & Wagener, T. 2012 When are multiobjective calibration trade-offs in hydrologic models meaningful? *Water Resources Research* **48** (3). https://doi.org/10.1029/2011WR011534.

Le Moine, N., Andréassian, V., Perrin, C. & Michel, C. 2007 How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water Resources Research* **43** (6). https://doi.org/10.1029/2006WR005608

Lu, J., Sun, G., McNulty, S. G. & Amatya, D. M. 2005 A comparison of six potential evapotranspiration methods for regional use in the Southeastern United States. *Journal of the American Water Resources Association* **41** (3), 621–633. https://doi.org/10.1111/j.1752-1688.2005.tb03759.x.

Nandakumar, N. & Mein, R. G. 1997 Uncertainty in rainfall-runoff model simulations and the implications for predicting the hydrologic effects of land-use change. *Journal of Hydrology* **192** (1–4), 211–232. https://doi.org/10.1016/S0022-1694(96)03106-X.

Omernik, J. M. 1987 Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* **77** (1), 118–125. https://doi.org/10.1111/j.1467-8306.1987.tb00149.x.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F. & Loumagne, C. 2005b Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *Journal of Hydrology. Elsevier* **303** (1–4), 290–306. doi:10.1016/J.JHYDROL.2004.08.026.

Oudin, L., Michel, C. & Anctil, F. 2005a Which potential evapotranspiration input for a lumped rainfall-runoff model?: Part 1—Can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs? *Journal of Hydrology* **303** (1–4), 275–289. doi:10.1016/J.JHYDROL.2004.08.025.

Parmele, L. H. 1972 Errors in output of hydrologic models due to errors in input potential evapotranspiration. *Water Resources Research* **8** (2), 348–359. https://doi.org/10.1029/WR008i002p00348.

Perrin, C., Michel, C. & Andréassian, V. 2001 Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology* **242** (3–4), 275–301. https://doi.org/10.1016/S0022-1694(00)00393-0.

Perrin, C., Michel, C. & Andréassian, V. 2003 Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology* **279** (1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7.

Roderick, M. L., Hobbins, M. T. & Farquhar, G. D. 2009a Pan evaporation trends and the terrestrial water balance. I. Principles and observations. *Geography Compass* **3** (2), 746–760. https://doi.org/10.1111/j.1749-8198.2008.00213.x.

Roderick, M. L., Hobbins, M. T. & Farquhar, G. D. 2009b Pan evaporation trends and the terrestrial water balance. II. Energy balance and interpretation. *Geography Compass* **3** (2), 761–780. https://doi.org/10.1111/j.1749-8198.2008.00214.x.

Samani, Z. A. & Pessarakli, M. 1986 Estimating potential crop evapotranspiration with minimum data in Arizona. *Transactions of the ASAE* **29** (2), 0522–0524. https://doi.org/10.13031/2013.30184.

Shahidian, S., Serralheiro, R., Serrano, J., Teixeira, J. L., Haie, N. & Santos, F. 2012 Hargreaves and other reduced-set methods for calculating evapotranspiration. *Evapotranspiration – Remote Sensing and Modeling* **23**, 50–80. https://doi.org/10.5772/18059.

Sivapalan, M., Yaeger, M. A., Harman, C. J., Xu, X. & Troch, P. A. 2011 Functional model of water balance variability at the catchment scale: 1. Evidence of hydrologic similarity and space-time symmetry. *Water Resources Research* **47** (2). https://doi.org/10.1029/2010WR009568

Smith, T., Marshall, L. & McGlynn, B. 2018 Typecasting catchments: classification, directionality, and the pursuit of universality. *Advances in Water Resources* **112**, 245–253. https://doi.org/10.1016/j.advwatres.2017.12.020.

Vaze, J., Jordan, P., Beecham, R., Frost, A., Summerell, G., Vaze, J., Jordan, P., Beecham, R., Frost, A. & Summerell, G. 2011 *Guidelines for Rainfall-Runoff Modelling Towards Best Practice Model Application*. eWater Cooperative Research Center. https://doi.org/978-1-921543-51-7.

Vázquez, R. F. 2003 Effect of potential evapotranspiration estimates on effective parameters and performance of the MIKE SHE-code applied to a medium-size catchment. *Journal of Hydrology* **270** (3–4), 309–327. https://doi.org/10.1016/S0022-1694(02)00308-6.

Verstraeten, W., Veroustraete, F. & Feyen, J. 2008 Assessment of evapotranspiration and soil moisture content across different scales of observation. *Sensors* **8** (1), 70–117. https://doi.org/10.3390/s8010070.

Vörösmarty, C. J., Federer, C. A. & Schloss, A. L. 1998 Potential evaporation functions compared on US watersheds: possible implications for global-scale water balance and terrestrial ecosystem modeling. *Journal of Hydrology* **207** (3–4), 147–169. https://doi.org/10.1016/S0022-1694(98)00109-7.

Wang, D. & Hejazi, M. 2011 Quantifying the relative contribution of the climate and direct human impacts on mean annual streamflow in the contiguous United States. *Water Resources Research* **47** (9). https://doi.org/10.1029/2010WR010283.

Xu, C. Y. & Singh, V. P. 2002 Cross comparison of empirical equations for calculating potential evapotranspiration with data from Switzerland. *Water Resources Management* **16** (3), 197–219. https://doi.org/10.1023/A:1020282515975.

First received 13 May 2020; accepted in revised form 30 November 2020. Available online 28 December 2020