Evaluation of Multi- and Many-Objective Optimization Techniques to Improve the Performance of a Hydrologic Model Using Evapotranspiration Remote-Sensing Data

Matthew R. Herman[1]; J. Sebastian Hernandez-Suarez[2]; A. Pouyan Nejadhashemi[3]; Ian Kropp[4]; and Ali M. Sadeghi[5]

Abstract: In this study, we explore the use of different multi- and many-objective calibration approaches in hydrological modeling when considering both observed streamflow and remotely sensed actual evapotranspiration (ETa). Eight remotely sensed ETa and an Ensemble products were used in a watershed in Michigan. Regarding the calibration process, the Unified-Non-dominated Sorting Genetic Algorithm III was integrated with the soil and water assessment tool (SWAT). The first nine calibrations used a multi-objective approach with two variables, one being streamflow and the other being a remotely sensed ETa products/Ensemble. The tenth calibration was a many-objective calibration with nine objective functions that represented observed streamflow and all eight of the remotely sensed evapotranspiration datasets. Results showed that the multi-objective calibrations were able to successfully calibrate both streamflow and ETa. However, the highest model performances were achieved using the Ensemble ETa product. Meanwhile, the required computational time for the many-objective calibration is significantly higher than the multi-objective calibration. In addition, the overall performance of many-objective method can be improved by considering weighting factors and constraining the search space. **DOI: 10.1061/(ASCE)HE.1943-5584.0001896.** © *2020 American Society of Civil Engineers.*

Author keywords: Actual evapotranspiration; Unified non-dominated sorting genetic algorithm III (U-NSGA-III); Ensemble; Hydrological model; Soil and water assessment tool (SWAT).

Introduction

Unchecked anthropogenic activities have led to the degradation of natural systems that are vital to society and life. In particular, the impacts on the Earth's limited freshwater supplies in combination with the increasing demand have made freshwater monitoring a major focus for researchers worldwide (Gleick 1993; Srinivasan et al. 2017; Haddeland et al. 2014). This requires the collection of data describing how different components of the hydrological cycle change across space and time. Traditionally, this has been accomplished by using land-based monitoring stations, which are able to collect highly accurate measurements of different components of the hydrological cycle (USGS 2018). However, monitoring stations are often expensive to install and maintain and are unable to provide the spatial resolution needed for large-scale analysis (Wanders et al. 2014). This has led to the introduction and use of hydrological models (Einheuser et al. 2013). These models are fast, inexpensive, and versatile tools compared with monitoring stations. However, because no model can perfectly characterize all processes within a watershed, a level of uncertainty is associated with all modeling practices (Kusre et al. 2010).

One way to improve model performance is through the calibration and validation process. However, because of the complex nature of the hydrological cycle, hydrological models use hundreds of parameters to describe the natural world. Each parameter has a default value assigned by the model based on watershed attributes, experimental results, or existing estimations for other areas of study. However, default values often do not accurately represent real-world conditions; therefore, these values need to be adjusted to improve model performance (Rajib et al. 2016). This is accomplished by modifying the parameters' values and comparing the model output to observed data. In hydrological modeling, this is traditionally done by comparing simulated and observed streamflows with statistical criteria to test model performance (Wanders et al. 2014). However, since hydrological models are used to simulate several state and flux variables of the hydrological cycle, using just one variable in the calibration process could result in poor performances in other hydrological components, which reduces the overall model performance while increasing predictive uncertainty (Wanders et al. 2014; Rajib et al. 2016). Therefore, it is important to consider several hydrological variables in the model calibration process (Crow et al. 2003). Among these variables, evapotranspiration (ET) stands out as an ideal addition to model calibration because it describes the loss of water from plants and the Earth's surface, which in turn drives weather patterns (Pan et al. 2015). However, land-based ET measurements are not widely available at large scales.

[1]Graduate Student, Dept. of Biosystems and Agricultural Engineering, Michigan State Univ., East Lansing, MI 48824.

[2]Graduate Student, Dept. of Biosystems and Agricultural Engineering, Michigan State Univ., East Lansing, MI 48824. ORCID: https://orcid.org/0000-0001-7155-9520

[3]Michigan State University Foundation Professor, Dept. of Biosystems and Agricultural Engineering, Michigan State Univ., East Lansing, MI 48824 (corresponding author). Email: pouyan@msu.edu

[4]Graduate Student, Dept. of Biosystems and Agricultural Engineering, Michigan State Univ., East Lansing, MI 48824. ORCID: https://orcid.org/0000-0002-3779-2801

[5]Retired, 11800 Sweet Land Way, Columbia, MD 21044; formerly, Research Physicist, Hydrology and Remote Sensing Laboratory, US Dept. of Agriculture-Agricultural Research Service, Beltsville, MD 20705.

Note. This manuscript was submitted on June 11, 2019; approved on October 11, 2019; published online on February 13, 2020. Discussion period open until July 13, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydrologic Engineering***, © ASCE, ISSN 1084-0699.

One solution to this is remotely sensed products. Remote sensing is the use of sensors and imaging equipment to indirectly measure the characteristics of an object (Graham 1999). When coupled with satellite technology, remote sensing has resulted in the development of various global monitoring datasets that can be used to measure elements of the hydrological cycle (Long et al. 2014). In particular, remote sensing has become a source of monitoring data for actual evapotranspiration (ETa), which describes the actual loss of water from both evaporation and transpiration (USGS 2016a). Methods for estimating ETa can be broadly classified into surface energy balance methods and physical-based methods. Surface energy balance methods employ land surface temperature (LST) from satellite thermal infrared imagery because LST is a signature of both ET and soil water availability (Stefan et al. 2015). Physical-based methods, on the other hand, estimate ETa as the sum of evaporation from the ground surface, plant transpiration, and evaporation from the precipitation intercepted by the canopy. These methods usually involve the application of the Penman-Monteith approach and soil water balances. Thus, input data can include land-based or assimilated meteorological observations in addition to optical remote sensing datasets. A known drawback of physical-based methods is their limited representation of secondary water sources and sinks such as irrigation, groundwater extraction, lateral redistribution of water, and tile drainage, among others, which affect evapotranspiration estimates (Hain et al. 2015). In contrast, surface energy balance models, based on remotely sensed LST, may account for the impact of non-precipitation sources and anthropogenic sinks (Hain et al. 2015). Still, physical-based methods such as land surface models have more flexibility in providing continuous ETa estimates over time, whereas LST-based products heavily rely on satellite's overpass frequency.

Regarding surface energy balance methods, a variety of remotely sensed ETa products have already been developed, including the Operational Simplified Surface Energy Balance (SSEBop) (Senay et al. 2013), the Atmosphere-Land Exchange Inverse (ALEXI) (Anderson et al. 2007), and the Google Earth Engine Evapotranspiration Flux (Google 2018). Meanwhile, available ETa products derived using physical-based methods include the Moderate Resolution Imaging Spectroradiometer (MODIS) Global Evapotranspiration Project (MOD16) (NTSG 2018), the North American Land Data Assimilation Systems phase 2 (NLDAS-2) (Xia et al. 2015), and the recently developed Terraclimate global ET dataset (Abatzoglou et al. 2018). Each of these products has different inputs, temporal and spatial resolutions, and methodologies for estimating ETa. Thus, it is challenging to select the most suitable dataset for calibrating/validating a watershed model. Furthermore, it is important that, although remotely sensing helps with the issue with improving data quantity and availability, it does not directly solve the issue of data quality. One way to mitigate the uncertainty associated with remotely sensed products is the use of ensemble techniques, which aim to combine the benefits of each product while accounting for their limitations (Dietterich 2000; Duan et al. 2007). Here again, different techniques have been developed, ranging from very simple calculations to complex modeling approaches (Lee et al. 2017; Wang et al. 2018). Furthermore, some techniques require the use of accurate observed data to determine which remotely sensed products are more accurate for the region of study (Kim et al. 2015).

The use of both ET and streamflow datasets in hydrological model calibration has been the focus of recent research. In an early work, Zhang et al. (2009) used the leaf area index (LAI) derived from MODIS/TERRA and the Penman-Monteith equation for estimating ETa. Then, a lumped conceptual daily rainfall-runoff model was calibrated using both streamflow observations and the MODIS-LAI derived ETa. Results indicated that model calibration using both variables produced better simulations of daily and monthly runoff compared with model calibration using only streamflow data. More recently, Rientjes et al. (2013) calibrated the conceptual HBV rainfall-runoff model using daily ETa estimates derived from MODIS/TERRA satellite images with the surface energy balance system. Results showed an overall better reproduction of water balance when simultaneously using both streamflow and ET datasets during model calibration. Likewise, Vervoort et al. (2014) used MODIS 16A3 ETa to calibrate a low parameter conceptual model and assess its performance in ungauged watersheds. However, results showed that streamflow performance was adversely affected when considering the ETa dataset for model calibration. This was mainly attributed to structural inadequacies in the conceptual model. In these studies, ETa time series were averaged at the watershed scale, and single, aggregated objective functions considering performance in both streamflow and ETa were used for model calibration. Meanwhile, regarding distributed and semidistributed hydrological models, ET datasets have been used for providing additional constraints during model calibration while reducing equifinality. For instance, Tobin and Bennett (2017) constrained ET-sensitive parameter values in the Soil and Water Assessment Tool (SWAT) by using the Global Land Evaporation: the Amsterdam Model dataset. As a result, hydrograph's recession curve representation was improved, although overall performance did not increase. Similarly, Wambura et al. (2018) constrained ET-related SWAT model parameters by identifying spatial patterns in multi-temporal MODIS ET datasets. In both previously mentioned studies, ET and streamflow were handled separately on a stepwise basis during the calibration process.

Most of the previous studies did not consider the calibration of parameters related to vegetation dynamics, energy utilization, and land-atmosphere interaction (i.e., biophysical parameters), rather focusing on other components such as runoff, subsurface flow, and river routing (Rajib et al. 2018). Therefore, some results have reported improvements in ETa representation but limited the influence on streamflow simulation when using remotely sensed ETa datasets during model calibration (Rajib et al. 2018). Nevertheless, there are cases that have explicitly incorporated biophysical parameters into multi-objective model calibration. For instance, Herman et al. (2018) obtained Pareto fronts describing the tradeoffs between streamflow and ETa representations by separately calibrating three vegetation-based model parameters affecting ETa. In that study, an evolutionary multi-objective optimization algorithm was used and an aggregated remotely sensed ETa monthly time series for the entire area of study was used as a reference during model calibration. However, the exclusion of streamflow-related model parameters during the multi-objective model calibration might have hindered parameter interactions capable of further improving overall model performance. Meanwhile, Rajib et al. (2018) evaluated different calibration configurations to quantify the effects of remotely sensed ETa data on streamflow and ETa simulations. Multi-objective calibration configurations in that study consisted of minimizing a single, aggregated objective function for both streamflow and ETa. The results of that study indicated that calibrating each subbasin individually provided better overall model accuracy and reduced uncertainty in water yield simulation. However, it was suggested that a suboptimal condition was achieved when considering a lumped set of calibration parameters for the entire study area because of limitations of the approach used (i.e., objective function aggregation) in handling multiple objectives simultaneously. In addition, the previously mentioned issue might be also attained when calibrating individual subbasins.

Generally, a wide range of techniques and remotely sensed products are available for hydrologic model calibration. The use of a single-objective approach, based on the aggregation of individual objective functions, limits the understanding of how several sources of information contribute to overall model performance. Moreover, there is a lack of study on comparison among ETa remotely sensed products and the role of different multi-objective (i.e., two to three objective functions) and many-objective (more than three objective functions) calibration approaches on the improvement of hydrologic model performance when considering both traditional and biophysical calibration parameters, which is the goal of this study. To achieve this goal several objectives will be explored including: (1) evaluation of the performance of a hydrologic model using a multi-objective calibration approach and individual ETa products, and (2) evaluation of the performance of a hydrologic model using a many-objective calibration approach and combined ETa products.

Methodology

Study Area

For this study, the Honeyoey Creek-Pine Creek Watershed (Hydrologic Unit Code 0408020203) located about the middle of the Lower Peninsula of Michigan was used to evaluate the applicability of remote-sensing products to improve the overall performance of hydrologic models (Fig. 1). This watershed is a part of the Saginaw Bay Watershed, which has a final outlet to Lake Huron. Covering approximately 1,100 km^2, the region is predominantly used for agriculture, with about 52% of the land devoted for crop production. After agriculture, the next major land covers are forests (~23%), wetlands (~17%), pasturelands (~5%), and finally, urban (~3%).

This region was selected for testing the remote sensing products because there is a lack of land-based spatial monitoring data in the area that can be used to setup a hydrologic model. In fact, the closest source of observed ETa data is the AmeriFlux stations located

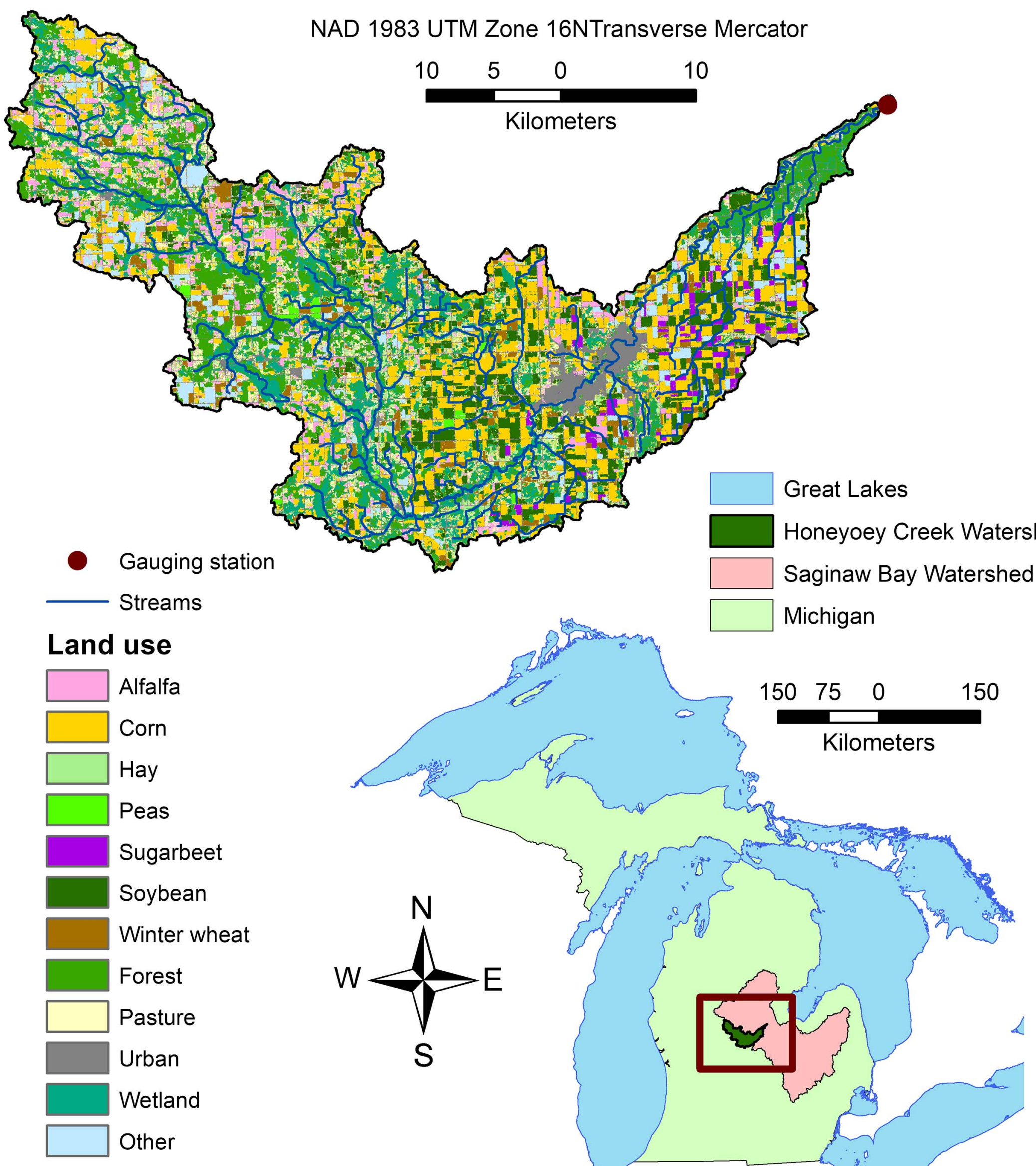

Fig. 1. Study area location and major land uses.

about 116 km away from the Honeyoey watershed (AmeriFlux 2018). All of this shows that remote sensing could serve a vital role in this region by providing consistent datasets for monitoring of the hydrological conditions. Meanwhile, streamflow is monitored on a daily basis at the outlet of the watershed by United States Geological Survey (USGS) station (USGS 2016b) and the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center (NCDC) has four stations in the region that measure daily precipitation and temperature (NCDC 2018).

Hydrological Model

To evaluate the hydrological cycle in the Honeyoey watershed, SWAT was selected. SWAT is a commonly used hydrological model that is time continuous and semi-distributed and was developed by the USDA Agricultural Research Service and Texas A&M AgriLife Research (Texas A&M University 2017). SWAT is able to simulate a variety of different hydrological processes and scenarios by taking into account the physiographic and climatological characteristics of the region of study (Gassman et al. 2007).

Relevant to this study is the way in which SWAT simulates ETa. However, before calculating ETa, the potential evapotranspiration should be estimated. In SWAT, this can be accomplished with three different techniques: (1) Penman-Monteith, (2) Priestley-Taylor, and (3) Hargreaves. The Penman-Monteith method, which is the default choice in SWAT, was used in this study. After calculating potential evapotranspiration, SWAT takes into account evaporation and transpiration from several sources including the evaporation from rainfall intercepted by the canopy, maximum transpiration, maximum soil evaporation, and sublimation (during periods of snow cover) (Neitsch et al. 2011). These are calculated at each hydrologic response unit (HRU), which are geographical units with homogenous land use, soil type, and slope characteristics. In this study, the Honeyoey watershed was divided into 250 subbasins, each one consisting of a single HRU (i.e., no spatial variability within the subbasins was considered). By calculating the potential evapotranspiration and taking into account the sources of evaporation and transpiration at the subbasin level, SWAT is able to report monthly ETa values for the entire region.

To develop the SWAT model for the Honeyoey watershed, several spatial and temporal datasets were used. They included topography, landuse, soil characteristics, climatological conditions, and crop management practices. For regional topography, the 30 m National Elevation Dataset from the USGS was used to calculate the watershed slope (NED 2014). Meanwhile, 30 m landuse data was obtained from the 2012 Cropland Data Layer, which was developed by the United States Department of Agriculture (USDA) – National Agricultural Statistics Service (NASS 2012). Regional soil characteristics were obtained on a scale of 1:250,000 from the Natural Resources Conservation Service (NRCS) Soil Survey Geographic Database (NRCS 2014). Climatological conditions for 2003–2014 were obtained separately from four NCDC stations (two temperature and two precipitation stations) (NCDC 2018). All other climatological conditions (e.g., wind speed, solar radiation, relative humidity) required for model simulation were generated using the WXGEN stochastic weather generator included in SWAT (Sharpley and Williams 1990; Wallis and Griffiths 1995; Neitsch et al. 2011). Crop management practices, which included operations, schedules, and crop rotations, were adopted from studies that used SWAT in the same region to account for local practices (Love and Nejadhashemi 2011). To perform the model calibration and validation, observed streamflow data was obtained from the Pine River USGS station located at the outlet of the Honeyoey watershed for the period from 2001 to 2014 (USGS 2016b). For this period, the first 2 years (2001–2002) were used for warming-up, and the next 12 years (2003–2014) were used for model calibration and validation.

Remote Sensing Actual Evapotranspiration Products

Given the lack of land-based ETa data in the region and to calibrate the SWAT model, eight different ETa products were obtained: (1) SSEBop, (2) ALEXI, (3) the MODIS Global Evapotranspiration Project (MOD16A2) 500 m, (4) the MOD16A2 1 km, (5) NLDAS-2 Mosaic, (6) NLDAS-2 Noah, (7) NLDAS-2 Variable Infiltration Capacity (VIC), and (8) TerraClimate. Each of these products uses different inputs and techniques to estimate ETa, also having different spatial and temporal resolutions. SSEBop and ALEXI products are obtained from surface energy balance models, MOD16A2 products were developed using a physical-based algorithm based on the Penman-Monteith approach, whereas the next four products (NLDAS-2 Mosaic, NLDAS-2 Mosaic Noah, NLDAS-2 VIC, and TerraClimate) were developed from land surface models. The following is a brief overview of each ETa product, its associated spatial and temporal resolutions, and overall performance. In addition, a summary of the main attributes of these datasets is provided in Table S1 (Supplementary Materials).

- The SSEBop ETa product developed by the USGS uses a simplified energy balance to calculate ETa on a monthly basis for the contiguous United States (Senay et al. 2013). This is accomplished by calculating ET fractions from 8-day, 1-km MODIS thermal imagery, which are then aggregated to a monthly scale (Senay et al. 2013; Velpuri et al. 2013). In general, SSEBop product underestimates ET over croplands, whereas MODIS shows higher performances for grassland and forest classes. In addition, there is not evidence of SSEBop product's accuracy changes with elevation (Velpuri et al. 2013). Despite its very simple formulation, SSEBop model is known for providing ETa estimates as good as more complex LST-based approaches such as METRIC (Mapping EvapoTranspiration at high Resolution with Internalized Calibration) (Singh and Senay 2016).
- The ALEXI was the product of a joint project between the USDA and the National Aeronautics and Space Administration (NASA) and is based on an energy balance. However, instead of ET fractions, ALEXI uses daily changes in surface temperature, obtained from Geostationary Operational Environmental Satellites, and relates those changes to surface water loss or ETa (Anderson et al. 2007). In addition, ALEXI has limited reliance on land-based meteorological inputs. The resulting product reports ETa on a daily time step at a 4-km spatial resolution for the contiguous United States. ALEXI has been widely validated for different climatic conditions and vegetation types, including rainfed and irrigated lands (Hain et al. 2015).
- The MOD16A2 500m and MOD16A2 1 km ETa products use the same methodology but have different landuse inputs and spatial resolutions (Running et al. 2018). These products were the result of a joint project between NASA and the University of Montana Numerical Terradynamic Simulation Group (NTSG 2018). ETa is calculated by using an algorithm that is based on the Penman-Monteith equation and also requires MODIS land cover, the fraction of photosynthetically active radiation/leaf area index, and the global surface meteorology (Mu et al. 2011; NASA 2018a, b). The results are global 8-day ETa products at 500 m or 1 km spatial resolutions depending on the inputs used (NASA 2014). Overall, the main uncertainty source for these products is landcover misclassification. In addition, MOD16 products are prone to underestimate cropland ETa and show higher accuracy in lower elevation zones. In general,

MOD16 ET results are comparable to SSEBop results (Singh and Senay 2016).
- The Mosaic, Noah, and VIC ETa products were developed as a joint project between the NOAA, the National Centers for Environmental Prediction (NCEP) Environmental Modeling Center, NASA's Goddard Space Flight Center, Princeton University, the University of Washington, NOAA's National Weather Service Office of Hydrological Development, and the NOAA/NCEP Climate Prediction Center and are part of the NLDAS project (NASA 2018c). Each of the products uses a different land surface model to take into account a variety of factors such as atmosphere interactions of water and energy, vegetation and soil moisture heterogeneity, water and energy budgets, and rainfall runoff and water storage (Xia et al. 2015). The ETa products that result from these models report ETa at both hourly and monthly time steps with a spatial resolution of 1/8 degree or 12 km for the entirety of North America (Long et al. 2014). Overall, Noah provides the most accurate ETa estimates when comparing against in situ observations (i.e., eddy covariance flux towers), followed by VIC and Mosaic, respectively. Particularly, the Mosaic model tends to overestimate ETa. In addition, performance is generally better for spring and fall seasons, and for deciduous forest and grassland vegetation types. However, Noah greatly underestimates ETa in spring and early summer, which is explained by a model structure limitation (Xia et al. 2015). The main uncertainty sources rely on meteorological input data, model calibration, and model structure deficiencies (Xia et al. 2015).
- The TerraClimate ETa product used a water-balance model and was the result of a joint project between the University of Idaho, the University of Montana, and the USDA Forest Service – Rocky Mountain Research Station. The water-balance model used by TerraClimate is based on the one-dimensional modified Thornthwaite-Mather climatic model (Abatzoglou et al. 2018). The output of this mode is a global ETa product that has a monthly time step with a 4-km spatial resolution (Abatzoglou et al. 2018). TerraClimate has been successfully validated against a variety of ETa networks, including the Global Historical Climate Network, SNOTEL, and RAWS. Meanwhile, several limitations were reported with this dataset such as lack of validation in regions with low observed value densities (Climatology Lab 2018). Also, the accuracy of this product is low in arid regions because of the simplicity of the water balance model used in TerraClimate (Zhao et al. 2019).

As discussed in the Introduction, although remote sensing provides access to global spatially distributed datasets, it also has more uncertainty associated with it. To minimize the level of uncertainty, the ETa products used in this study went through an extensive calibration and validation process against ground-based observations. The results of these analyses were reported in the Supplementary Materials (Table S1). However, these accuracies are based on comparisons to specific locations where observed data was available. Given the nature of ETa, these accuracies could flux across the landscape. This means that for each dataset it may perform better, equal, or worse in any other locations such as the Honeyoey watershed. However, the goal of this study is not to perform revalidation but to explore modeling applications of remote sensing ETa products to improve the performance of physically based hydrologic models in regions lacking observed data. This means that while the model calibrations cannot confirm the best ETa product to use globally, it can highlight those that perform better than others in a region. This can be measured by comparing the level of improvement in the model predictability of streamflow using different ETa products since observed streamflow data are more available than observed ETa.

Nevertheless, an ensemble of the ETa products is also used in this study for the model calibration to help to reduce the uncertainty level associated with the ETa products. Concerning ensemble techniques for ETa products, Bayesian Model Averaging (BMA) is commonly used (Kim et al. 2015; Tian and Medina 2017; Yao et al. 2017; Ma et al. 2018). BMA is the preferred method as it reduces overall uncertainty by determining weights for each ETa product by comparing them to observed data (Kim et al. 2015). However, when the observed data are not available, an averaging technique can be used (Tebaldi and Knutti 2007). In this study, the arithmetic mean method was used due to a lack of observed ETa data in the study area.

Calibration Techniques

We implemented a multi-objective calibration approach in order to account for multiple sources of information describing both streamflow and ETa variables. The overall calibration process consisted of (1) processing remote sensing products to obtain monthly ETa time series for the entire study area, (2) selecting model calibration parameters, (3) defining objective functions for each variable, (4) formulating the multi-objective optimization problem to solve, (5) selecting and implementing suitable multi- and many-objective optimization algorithms, and (6) selecting the best tradeoff solution for analysis purposes.

For this study, we implemented two different strategies in the formulation of the multi-objective calibration process. In the first strategy, we implemented a multi-objective approach in which the performance of a hydrological model was evaluated using individual ETa remote sensing products. Meanwhile, in the second strategy, we used a many-objective approach to evaluate the performance of a hydrological model when all ETa remote sensing products were simultaneously used. For both strategies, we opted for implementing multi-objective optimization approaches, avoiding the subjective formulation of a single aggregated objective function. In addition, the multi- and many-objective frameworks provide a set of optimal solutions describing the tradeoffs between streamflow and ETa performances while incorporating external sources of information describing both variables. In this study, we employed the recently proposed evolutionary optimization algorithm – Unified Non-dominated Sorting Genetic Algorithm III (U-NSGA-III) (Seada and Deb 2016), which is capable of solving different types of problems (i.e., single-, multi-, and many-objective). Furthermore, to compare the resulting Pareto-optimal solutions, we selected the best tradeoff solution by using the compromise programming approach. A detailed description of the calibration approach is presented in the following section.

Data Processing

During the calibration process, the observed data consisted of eight raster-based remote sensing products for ETa, and a unique observed streamflow daily time series from a USGS gauging station located at the watershed's outlet. To perform multi- and many-objective calibrations, both streamflow and ETa products should be represented at the watershed outlet. Therefore, weighted averaged ETa values were used here along with streamflow. The times series used for both streamflow and ETa were obtained for the period 2003–2014.

Calibration Parameters

Because the goal of the study is to improve the simulation skills of the SWAT model for ETa and streamflow, we identified the main parameters that are relevant to the ETa and streamflow processes for calibrating the model. Based on the literature review, SWAT documentation, and sensitivity analysis, 18 parameters were selected for this study (Arnold et al. 2012; Woznicki and Nejadhashemi 2012). These parameters are: baseflow recession constant (ALPHA_BF), biological mixing efficiency (BIOMIX), maximum canopy storage (CANMX), effective hydraulic conductivity of channel (CH_K2), Manning's n value for the main channel (CH_N2), moisture condition II curve number (CN2), plant uptake compensation factor (EPCO), soil evaporation compensation coefficient (ESCO), fraction of maximum stomatal conductance corresponding to the second point on the stomatal conductance curve (FRGMAX), maximum stomatal conductance at high solar radiation and low vapor pressure deficit (GSI), delay time for aquifer recharge (GW_DELAY), revap coefficient (GW_REAP), threshold water level in shallow aquifer for base flow (GWQMN), aquifer percolation coefficient (RCHRG_DP), threshold water level in shallow aquifer for revap (REVAPMN), available water capacity (SOL_AWC), surface runoff lag coefficient (SURLAG), and the vapor pressure deficit corresponding to the second point on the stomatal conductance curve (VPDFR). The minimum, maximum, and default values for all these parameters are presented in Table 1.

Objective Functions

For each variable, we formulated a minimization objective function (f) based on the Nash-Sutcliffe efficiency NSE, as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \tag{1}$$

$$f = 1 - NSE \tag{2}$$

where O_i is the ith observation of the considered variable (i.e., streamflow or ETa); $\bar{O}$ is the average of the observed data; P_i is the ith simulated value of the considered variable; and n is the total number of observations. The range of the resulting objective functions spans from zero to infinity, where zero represents a perfect fit between simulated and observed time series. We used daily time series for streamflow, whereas for ETa, we used a monthly time step. The monthly ETa time series obtained for each remote sensing product was considered as the observed ETa data.

Table 1. SWAT parameters considered during the model calibration process

Parameter	Minimum value	Maximum value	Default value
ALPHA_BF	0	1	0.048
BIOMIX	0	1	0.2
CANMX	0	100	0
CH_K2	0	500	0
CH_N2	0	0.3	0.014
CN2[a]	−25%	25%	Various
EPCO	0	1	1
ESCO	0	1	0.95
FRGMAX	0	1	Various
GSI	0.001	0.05	Various
GW_DELAY	0	500	31
GW_REVAP	0.02	0.2	0.02
GWQMN	0	5,000	1,000
RCHRG_DP	0	1	0.05
REVAPMN	0	1,000	750
SOL_AWC[a]	−25%	25%	Various
SURLAG	1	24	4
VPDFR	1.5	6	Various

[a]These parameters are treated as global multiplying factors that modify the assigned values for each HRU.

For each simulation in the optimization process, we computed only one objective function for streamflow using the available observed dataset at the outlet of the study area, and as many objective functions for ETa as the number of ETa datasets.

Optimization Strategies

We implemented two calibration strategies to evaluate the influence of the different ETa datasets in the prediction of daily streamflows. In the first strategy (multi-objective optimization), we formulated several bi-objective optimization problems to simultaneously minimize the difference between observed and simulated time series for both streamflow and ETa. For each bi-objective problem, we used a different ETa dataset. Moreover, in this strategy, we formulated an additional bi-objective optimization problem using an ensemble ETa dataset. This ensemble dataset was computed by averaging the monthly values of the ETa time series from each individual remote sensing product. As a result, for the first strategy, we obtained as many Pareto-optimal fronts as the number of ETa datasets used in this study. Each optimization problem was formulated as follows:

$$\min_{\theta \in \Omega} F(\theta) = [f_Q(\theta), f_{ET}(\theta)] \tag{3}$$

where F is a vector composed of multiple objective functions; θ is a vector containing values for p model calibration parameters; Ω is a p-dimensional parameter space limited by the calibration ranges for each model parameter; f_Q is the objective function evaluated for streamflow; and f_{ET} is the objective function evaluated for ETa.

In the second strategy (many-objective optimization), we simultaneously minimized all the objective functions derived from each ETa datasets and for the streamflow variable. In this strategy, we did not include the ensemble dataset to avoid the addition of redundant information into the overall optimization process. Hence, this strategy results in one Pareto-optimal front. The many-objective optimization problem was formulated as follows:

$$\min_{\theta \in \Omega} F(\theta) = [f_Q(\theta), f_{ET_1}(\theta), f_{ET_2}(\theta), \ldots, f_{ET_S}(\theta)] \tag{4}$$

where f_{ET_s} is the objective function evaluated using the sth ETa dataset and S is the total number of ETa datasets.

Multi-Objective Optimization Algorithm

The U-NSGA-III algorithm is an extension of the recently proposed NSGA-III algorithm (Deb and Jain 2014). The original NSGA-III is a population-based, elitist procedure based on reference directions, which uses non-domination sorting and evolutionary operators (i.e., crossover and mutation) to move toward an optimal Pareto front. Reference directions are vectors that evenly fill the objective space. The algorithm uses these vectors to rank the diversity of individuals (Deb and Jain 2014). Moreover, these vectors are normalized by default to achieve an equally diverse optimal Pareto front with respect to each objective function. NSGA-III's parameters are the population size, the number of generations, the crossover and mutation probabilities, and distribution indices associated with the crossover and mutation operations. This algorithm has been found to reduce its performance when working with two or one objective functions. However, by incorporating an explicit selection procedure when scaling down to two- and single-objective problems, the U-NSGA-III algorithm is capable of solving single-, multi- (i.e., two to three objective functions), and many-objective (i.e., more than three objective functions) optimization problems without adding extra parameters (Seada and Deb 2016).

Prioritizing streamflow and ETa for the many-objective optimization strategy described in the previous section, posed an interesting challenge. By default, U-NSGA-III equally prioritize all the objectives. Therefore, ETa calibration holds most of the total weight of the overall search for this strategy because here we have eight ETa remote sensing products. Likewise, optimization under these default settings may result in a poor calibration performance for streamflow. To improve performance, the balance of weights along each objective was modified by manipulating reference direction vectors to award the same amount of weight to the streamflow calibration objective function as all the functions together.

To modify the weights, the set of $(M+1)$-dimensional reference directions is simplified into a two-dimensional reference direction set. The first dimension represents the weight given to all of the M ETa objectives, and the second dimension represents the weight given to the streamflow objective. The reference directions are created by the Das-Dennis method that generates normalized reference direction vectors (Das and Dennis 1998). At this stage, the reference directions can be represented in the following matrix:

$$\begin{bmatrix} r_{11} & r_{12} \\ \vdots & \vdots \\ r_{d1} & r_{d2} \end{bmatrix}$$

where r_{xy} is the yth dimension for the xth reference direction and d is the number of reference directions to create. Because the objectives are equally weighted according to the Das-Dennis method, the following relationship must be attained:

$$(r_{x1} + r_{x2} = 1) \quad \forall\ x \tag{5}$$

Then, we generate a $d \times (M+1)$ reference direction matrix out of the $d \times 2$ matrix by splitting the first dimension (representing ETa) into M different weights for each of the ETa objective functions:

$$\begin{bmatrix} r_{11}w_1 & r_{11}w_2 & \dots & r_{11}w_M & r_{12} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{d1}w_1 & r_{d1}w_2 & \dots & r_{d1}w_M & r_{d2} \end{bmatrix}$$

where w_z is zth weight for each of the ETa objectives, such that $w_z \in \mathbb{R},\ 0 \le w_z \le 1$. Because all ETa datasets will be equally weighted (i.e., $w_z = 1/M \forall z$) and, because $\sum_{z=1}^{M} w_z = 1$, the sum of the elements of each row in the $d \times (M+1)$ matrix showed above must be equal to 1:

$$\left[\left(\sum_{z=1}^{M} w_z r_{x1}\right) + r_{x2} = 1\right] \quad \forall\ x \tag{6}$$

Therefore, here we developed d $(M+1)$-dimensional reference directions, where the first M dimensions together have the same weight as the $(M+1)$th dimension. These reference directions were then used in U-NSGA-III to provide the same weight for the M objective functions for ETa as for the streamflow objective function.

The next step is to determine the number of reference directions and population members. The chosen method for calculating the number of reference directions, as described in Deb and Jain (2014), is the Das-Dennis method, which is popular among NSGA-III implementations. In Deb and Jain (2014), problems below eight objectives have the following number of reference directions H:

$$H = \binom{M+q-1}{q}$$

where M represents the number of objectives and q is a parameter in the Das-Dennis method. Higher q values create a denser set of reference directions. Thus, different q values for different M values are recommended (Deb and Jain 2014). Because the number of reference directions grows very large for $M > 7$, Deb and Jain (2014) developed a new approach for determining H. Instead of creating a single large set of reference directions; the algorithm creates two nested and less dense sets of reference directions. The outer layer has a q of 3, and the inner layer has a q of 2 and is half as large. This method results in 210 reference directions. Once the number of reference directions is known, we can determine the number of population members. Traditionally, the population is the lowest number divisible by four above the number of reference directions. In our case, that would be 212 population members. Finally, we also set a constraint for the streamflow objective function, where only NSE values greater than 0 are considered feasible.

The Computational Optimization and Innovation Laboratory at Michigan State University provided the java code for the U-NSGA-III algorithm. This code was adapted for calibrating the SWAT model for this study. To the best of our knowledge, this is the first time that the U-NSGA-III algorithm is used for hydrologic model calibration using both streamflow and ETa remotely sensed products.

Best Tradeoff Solutions

To obtain the best tradeoff calibration solution considering both streamflow and ETa, in this study, we used the compromise programming approach using the l_2 metric (Zeleny and Cochrane 1973). This allows the selection of an individual point from each Pareto-optimal front attained after implementing either the multi- or many-objective optimization strategies. The metric was computed for each member i of the Pareto front as follows:

$$l_{2_i} = \sqrt{\sum_{j=1}^{M} \left(\frac{|f_{ij} - f_{ij}^{ideal}|}{|f_{ij}^{ideal} - f_{ij}^{nadir}|}\right)^2} \tag{7}$$

where j is an index identifying each objective function f; M is the number of objective functions; f^{ideal} is the vector containing the ideal point coordinates, which is an unfeasible solution located outside the Pareto front, representing the best-expected objective function values (in this case, zero); and f^{nadir} is the vector containing the nadir point coordinates, which comprises the worst objective function values obtained for each dimension at the optimal Pareto front. The point with the minimum l_2 metric (i.e., closest to the ideal point) is selected as the best tradeoff solution.

In addition to the best tradeoff solutions of the multiple Pareto fronts, we also calculated the worst expected solutions to obtain a common nadir point for all the bi-objective Pareto-optimal solutions. For this purpose, we identified the individual solution (i.e., model simulation) providing the minimum f_Q among all the Pareto fronts. Then, from this model solution, we computed the corresponding f_{ET} for each of the ETa datasets (i.e., we obtained the worst expected f_{ET} for each dataset). Thus, the nadir point was defined as the vector containing the maximum f_Q and the worst expected f_{ET} among all the Pareto-optimal solutions.

Calibration Evaluation

To determine how well the SWAT model replicated either the spatially distributed remotely sensed ETa or point measurement of streamflow records at the outlet of the watershed, three statistical criteria were used, which were recommended by Moriasi et al. (2007). The three criteria are: (1) Nash-Sutcliffe efficiency (NSE); (2) percent bias (PBIAS); and (3) the ratio of root-mean-square error (RMSE) to observed standard deviation ratio (RSR). NSE is a measure of the level of residual variance compared to actually

J. Hydrol. Eng., 2020, 25(4): 04020006

measured data variance (Nash and Sutcliffe 1970). PBIAS is a measure of the overall simulated dataset's tendency to be larger or smaller than the observed data (Gupta et al. 1999). RSR is, as its name implies, a ratio between the RMSE and the observed standard deviation (Singh et al. 2005). For a model to be satisfactorily calibrated the following criteria needed to be met: NSE > 0.5, PBIAS < ±25%, and RSR < 0.7 (Moriasi et al. 2007).

Results and Discussion

Evaluation of the Performance of the Hydrologic Model Using a Multi-Objective Calibration Approach and Individual ETa Products

The first goal of this study was to evaluate the improvement in the performance of a hydrological model in estimating the streamflow by comparing the potential benefit of using the eight different ETa products along with an Ensemble of all the datasets. To do this, nine SWAT models were calibrated by simultaneously adjusting 18 SWAT parameters affecting both streamflow and ETa estimation as it was discussed in the section "Calibration Parameters." Each calibration had two objective functions: (1) streamflow with observed data collected from a USGS station at the watershed outlet, and (2) ETa with each calibration using a different ETa product. This resulted in a total of 9 different calibrations, each of which had 38,600 simulations (i.e., 385 generations with a population size of 100), resulting in a total of 347,400 simulations across all bi-objective calibrations. The results of the U-NSGA-III calibrations are presented in Fig. 2 in the form of optimal Pareto frontiers. As can be seen, variability among the ETa products resulted in a wide range of performances. However, regarding individual models, the Ensemble ETa product showed the highest overall model performance while the TerraClimate ETa product showed the lowest overall model performance. This shows that the Ensemble, which was used to minimize the uncertainty associated with the individual ETa products, was successful and outperformed all products. However, it is also important to note that none of the ETa products were compared with land-based measurements within the study area, and thus the Ensemble cannot be labeled as the most accurate of the ETa products. Instead, the Ensemble was able to most closely replicate the SWAT model simulations. Meanwhile, because SWAT is a physically based model, the fact that the Ensemble outperformed all the other ETa products indicates that at least for this region, the Ensemble product is more aligned with current knowledge of water movement in a watershed.

To better understand Fig. 2 and how each calibration performed, a summary of each set of Pareto optimal solutions for each bi-objective optimization, with respect to the NSE statistical criteria, is presented in Table 2. This includes the mean, standard deviation, coefficient of variation, minimum, maximum, the best tradeoff solution, and the worst-case ETa performance. The maximum and minimum columns show the best and worst NSE values for each objective function. Here it is important to note that all cases fall within the satisfactory ranges for the NSE calibration criteria. This means that all of the potential solutions identified by the calibration process in the Pareto frontiers would be considered as acceptable models. However, the ETa model performance was higher than the streamflow performance, showing that the calibration process had a better fit in replicating the ETa products than the observed streamflow.

Meanwhile, in this study, the best or optimal solutions to each calibration are presented in the "Best Tradeoff" column. These cases were calculated as the solution that had the smallest distance to the origin (0, 0) in Fig. 2 for 1-NSE. This applied equal importance to streamflow and ETa. As can be seen, the tradeoff solutions do not achieve the same levels of model performance reported in the maximum column. For example, the maximum streamflow performance for the MOD16 1 km product is an NSE of 0.78, whereas the best tradeoff has a streamflow NSE of 0.77. Nevertheless, when comparing the individual calibrated model performances, the Ensemble had the best overall performance with a streamflow NSE of 0.79 and an ETa NSE of 0.95, whereas TerraClimate had the worst model performance with a streamflow NSE of 0.75 and an ETa NSE of 0.76. This shows that the Ensemble was able to

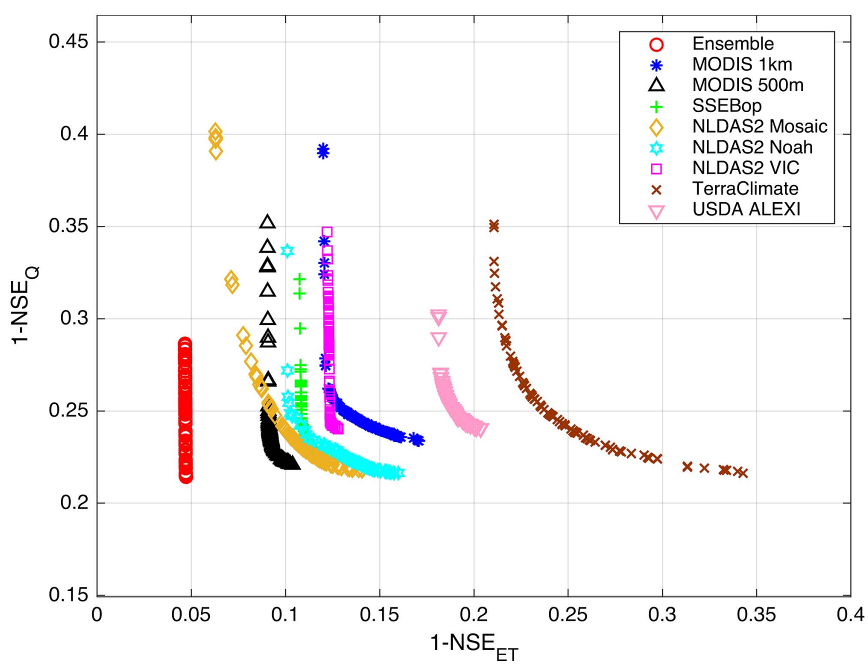

Fig. 2. Comparison of the Pareto frontiers of the nine multi-objective calibrated SWAT models.

Table 2. Summary of multi-objective calibration Pareto frontiers, where "Q" refers to streamflow performance and "ET" refers to actual evapotranspiration performance

Dataset	Mean		Standard deviation		Coefficient of variation (%)		Maximum		Minimum		Best tradeoff solution		Worst case
	NSE_Q	NSE_{ET}	NSE_Q	NSE_{ET}	NSE_Q	NSE_{ET}	NSE_Q	NSE_{ET}	NSE_Q	NSE_{ET}	NSE_Q	NSE_{ET}	NSE_{ET}
MOD16 500 m	0.76	0.91	0.027	0.003	3.5	0.4	0.78	0.91	0.65	0.90	0.78	0.90	0.89
MOD16 1 km	0.75	0.86	0.026	0.013	3.5	1.5	0.77	0.88	0.61	0.83	0.76	0.86	0.77
SSEBop	0.76	0.89	0.019	0.003	2.5	0.4	0.78	0.89	0.68	0.88	0.78	0.88	0.86
NLDAS2-Mosaic	0.76	0.89	0.040	0.018	5.3	2.1	0.78	0.94	0.60	0.86	0.78	0.88	0.85
NLDAS2-Noah	0.77	0.87	0.015	0.015	1.9	1.7	0.78	0.90	0.66	0.84	0.78	0.87	0.82
NLDAS2-VIC	0.72	0.88	0.029	0.001	4.1	0.1	0.76	0.88	0.65	0.87	0.76	0.87	0.83
TerraClimate	0.75	0.75	0.029	0.033	3.9	4.4	0.78	0.79	0.65	0.66	0.75	0.76	0.59
USDA-ALEXI	0.75	0.81	0.011	0.006	1.5	0.7	0.76	0.82	0.70	0.80	0.75	0.81	0.78
Ensemble	0.75	0.95	0.022	0.0001	2.9	0.01	0.79	0.95	0.71	0.95	0.79	0.95	0.95

outperform all of the individual ETa products, which also shows that using the Ensemble was successful at improving the overall model performance.

Meanwhile, by considering the standard deviation and coefficient of variation values, in general, ETa model performance had lower values than streamflow performance. However, when considering only streamflow model performance, the standard deviation ranged from 0.011 (ALEXI product) to 0.040 (NLDAS2-Mosaic product), and the coefficient of variation ranged from 1.5% (NLDAS2-Noah product) to 5.3% (NLDAS2-Mosaic product) (Table 2). This shows that during the calibration process, the NLDAS2-Mosaic product had the largest span in potential solutions, whereas the NLDAS2-Noah product had the smallest span. Meanwhile, when considering only the ETa model performance, the standard deviation ranged from 0.0001 (Ensemble product) to 0.033 (TerraClimate product), and the coefficient of variation ranged from 0.01% (Ensemble product) to 4.4% (TerraClimate product) (Table 2). This shows that during the calibration process the TerraClimate product had the largest span in potential solutions, whereas the Ensemble product had the smallest span.

Table 3 presents the performance metrics for the best tradeoff solutions identified for each bi-objective calibration. The reported values for NSE, PBIAS, and RSR indicate that both streamflow and ETa meet the criteria for good and satisfactory performances, respectively, according to Moriasi et al. (2007). In addition, Figs. 3 and 4 show graphical results for the best tradeoff solution obtained with the Ensemble dataset. In Fig. 3, the simulated time series are compared with the observed streamflow and the ETa dataset values. Meanwhile, Fig. 4 illustrates the relationship between observed and simulated values by using scatter plots. These scatter plots were obtained using both regular and log-transformed axes. Similar plots for the other eight ETa datasets are found in the Supplementary Data (Figs. S1–S25). In general, results indicate a satisfactory overall representation of the two simulated components of the hydrological cycle. However, low values for both streamflow and ETa are not very well represented. Particularly, the order of magnitude of the differences between observed and simulated values increases with a reduction in the streamflow magnitude [Fig. 4(c)]. Similarly, observed and simulated ETa values below ~10 mm/month show low correspondence [Fig. 4(d)]. In general, simulated ETa values are underestimated for low magnitudes, which may be a result of using NSE as an objective function that is very sensitive to high values (Legates and McCabe 1999).

Evaluation of the Performance of the Hydrologic Model Using a Many-Objective Calibration Approach and Combined ETa Products

The second goal of this study was to explore the novel use of a many-objective calibration approach for hydrological modeling. This was done by calibrating the SWAT model simultaneously using all eight ETa products and streamflow. In total, 252,704 simulations were run (1,191 generations with a population size of 212). We compared the results for the streamflow objective function for three scenarios: (1) equal weight for all nine objective functions with no constraint for the streamflow objective function, (2) balanced weights for all nine objective functions with no constraint for the streamflow condition, and (3) balanced weights for all nine objective functions with a constraint for the streamflow condition as described in section "Multi-Objective Optimization Algorithm." However, for the sake of argument, the results for the streamflow objective functions concerning the MOD16 500 m ETa objective function are presented in Fig. 5 for the previously mentioned scenarios.

Table 3. Performance metrics for the best tradeoff solutions obtained for the multi-objective calibration

Dataset	Streamflow				Actual evapotranspiration			
	NSE	PBIAS (%)	RSR	RMSE (m^3/s)	NSE	PBIAS (%)	RSR	RMSE (mm/month)
MOD16 500 m	0.78	9.56	0.47	4.93	0.90	9.41	0.32	13.05
MOD16 1 km	0.76	2.02	0.49	5.17	0.86	0.96	0.38	10.71
SSEBop	0.78	0.98	0.47	4.95	0.88	−19.74	0.35	14.38
NLDAS2-Mosaic	0.78	13.16	0.47	4.95	0.88	18.02	0.35	15.60
NLDAS2-Noah	0.78	2.86	0.47	4.95	0.87	−12.48	0.37	12.53
NLDAS2-VIC	0.76	6.09	0.49	5.13	0.87	−15.23	0.36	14.05
TerraClimate	0.75	12.08	0.50	5.22	0.76	6.58	0.49	18.75
USDA-ALEXI	0.75	13.01	0.50	5.20	0.81	14.63	0.44	15.46
Ensemble	0.79	6.21	0.46	4.85	0.95	1.82	0.22	7.98

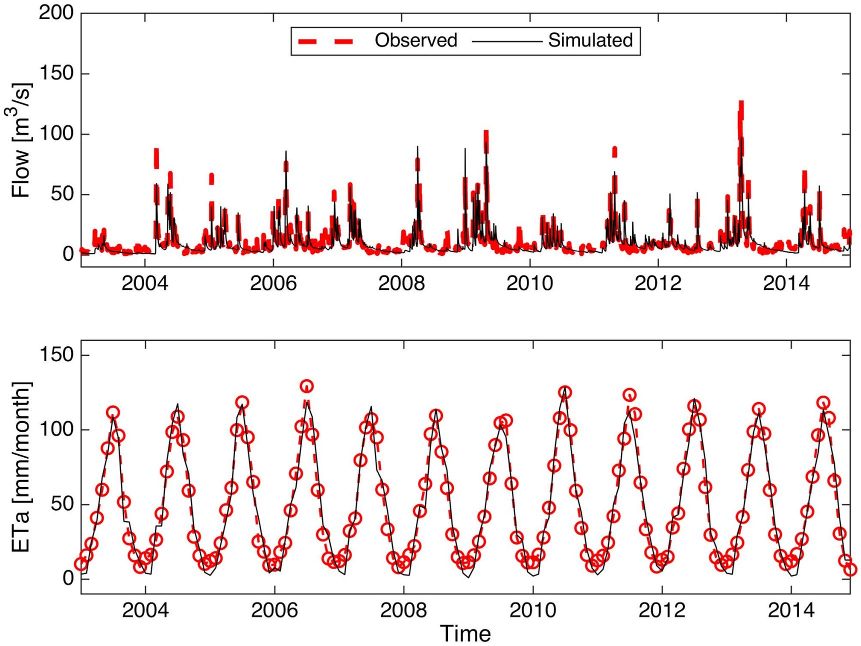

Fig. 3. Comparison of observed streamflow and remote-sensing-based ETa values with simulated time-series for the best tradeoff solution obtained with the ensemble ETa dataset.

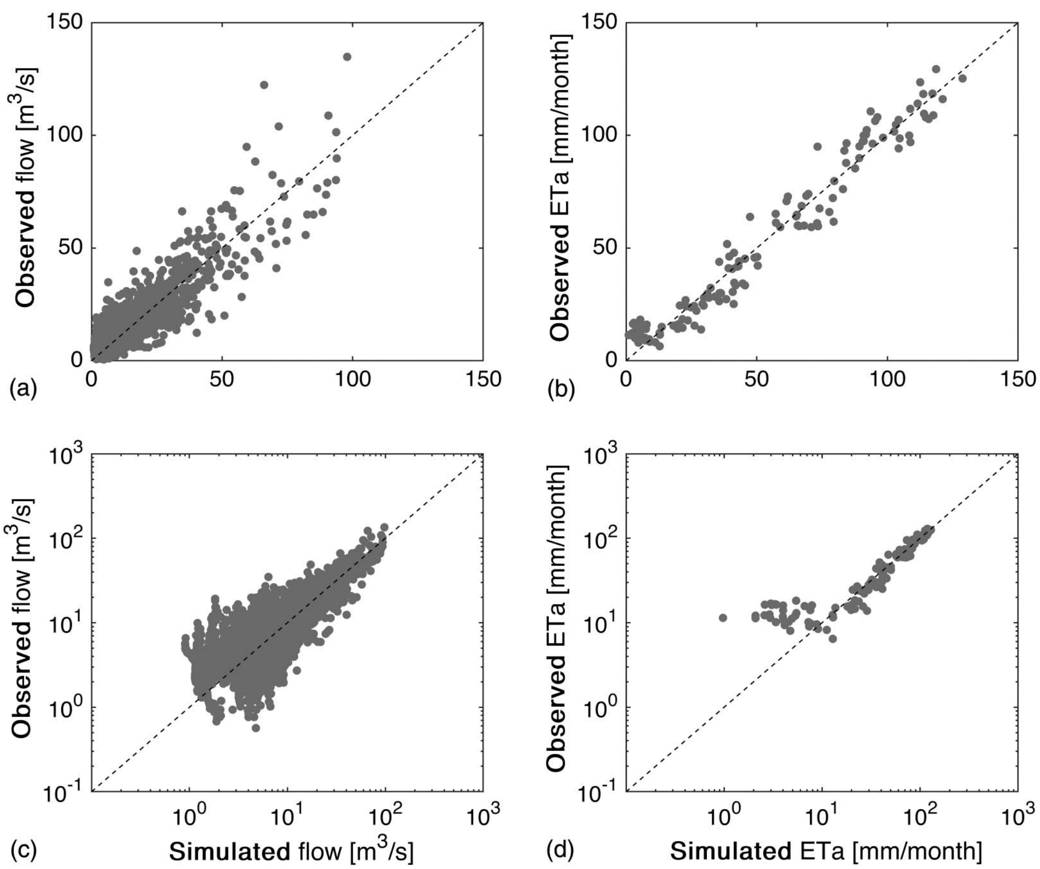

Fig. 4. Scatter plots comparing the observed and simulated values using the best tradeoff solution for the ensemble dataset: (a) streamflow; (b) ETa; (c) streamflow with log-transformed axes; and (d) ETa with log-transformed axis.

There is a vast difference between the results for the three calibration scenarios. For the equal weight calibration scenario [Fig. 5(a)], streamflow objective function values (i.e., $1 - \mathrm{NSE}_{\mathrm{streamflow}}$) vary from approximately 0.51–6.5. Thus, $\mathrm{NSE}_{\mathrm{streamflow}}$ ranges from −5.5 to 0.49, indicating that all Pareto-optimal solutions are outside the acceptable calibration range (NSE > 0.5) as described by Moriasi et al. (2007). This shows that the calibration was biased toward the eight ETa remote sensing products. Hence, the

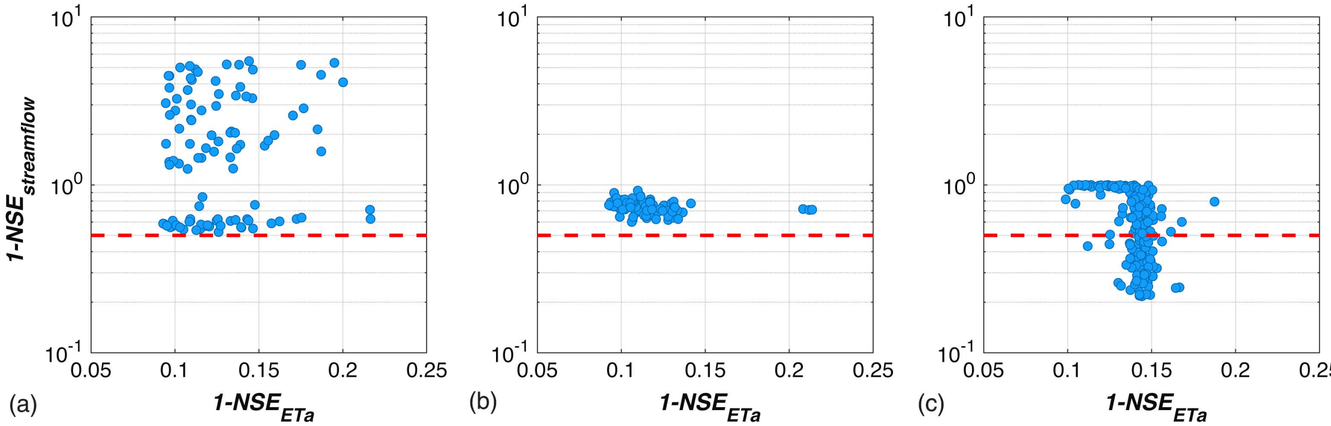

Fig. 5. Pairwise comparisons of the streamflow objective function and the ETa objective functions for the MOD16 500 m dataset for (a) many-objective calibration with equal weights; (b) many-objective calibration with balanced weights; and (c) many-objective calibration with balanced weights and constrained streamflow objective function. The dashed line represents the acceptability threshold (NSE = 0.5).

calibration was not successful for streamflow when using equal weights for all nine objective functions with no constraints. Meanwhile, the second many-objective calibration strategy [Fig. 5(b)] resulted in a much smaller range of $1 - \mathrm{NSE}_{\mathrm{streamflow}}$ values (0.5–2). This again shows a poor overall model calibration for streamflow, but the effect of balancing the weights for ETa objective functions shows a considerable improvement for the overall model performances. However, the problem with this approach is the very big search space created by a large number of objective functions. To narrow down the search space, in the third calibration scenario, a constraint was introduced for the streamflow objective function (i.e., $1 - \mathrm{NSE}_{\mathrm{streamflow}} < 1$). As a result, we were able to considerably improve the general performance of the model calibration for streamflow while maintaining good performances for ETa [Fig. 5(c)]. In this case, Pareto-optimal solutions resulted in $1 - \mathrm{NSE}_{\mathrm{streamflow}}$ values from approximately 0.2–0.8.

Meanwhile, Fig. 6 shows the Pareto-optimal values for ETa by using balanced weights and constrained streamflow objective functions (i.e., below 1). This figure indicates that ETa objective functions ranged from approximately 0.05–0.45. Similar results were obtained for the other two scenarios (i.e., many-objective calibration with equal weights and many-objective calibration with balanced weights). This indicates that all calibration runs were able to achieve satisfactory model calibration for ETa simulation. Furthermore, these results show that balancing the weights of the ETa objective functions and adding a constraint for the streamflow objective function, which improves the overall streamflow calibration, has little effect on ETa calibration.

Fig. 6 also includes Pearson's correlations values between all objective functions. These values are presented in the upper triangle. Considering the highly positive correlated datasets, it is possible to determine if the calibration process contains redundant datasets and then recommend a smaller set of ETa products to be used for model calibration. As can be seen, most of the objective functions have low positive correlation values ($r < 0.7$). However, there were two groups of ETa products that were highly correlated among each other. The first group includes the SSEBop, Noah, and VIC ETa products, whereas the second group includes the MOD16 500 m, ALEXI, and Mosaic products. The high correlations found among these products were also noticeable in the pairwise scatter plots (the lower triangle of Fig. 6). These high correlations are explained because these datasets are very similar in magnitude and pattern, resulting in similar objective functions when compared against simulated time series. Furthermore, the presence of these correlations indicates that a smaller set of ETa products could be used for SWAT model calibration. Here, the best tradeoff solution from the many-objective Pareto-optimal frontier can be identified by selecting the dataset with the highest NSE.

Table 4 shows the performance metrics for the best tradeoff solutions identified from the resulting many-objective Pareto frontier. Meanwhile, the model parameters associated with the best tradeoff solutions for the multi- and many-objective approaches are reported in Table S2. Using the information provided in Table 4, SSEBop was selected as a prefer dataset from the set comprised of the SSEBop, Noah, and VIC products. Likewise, MOD16 500 m would be the dataset to select from the set comprising the MOD16 500 m, ALEXI, and Mosaic products. Thus, the final ETa products set would include the MOD16 1 km, MOD16 500 m, SSEBop, and TerraClimate ETa products. Future studies should explore the use of this simplified ETa product set in hydrological model calibration. In addition, results indicate that streamflow attains a satisfactory performance (compared with the good performance attained for the multi-objective calibration), whereas ETa results show very good performance in terms of NSE and RSR. However, SSEBop and VIC datasets do not present satisfactory results in terms of PBIAS (<–25%). In summary, the many-objective calibration for the SWAT model is not as good as the multi-objective calibrations performed earlier, especially when using the Ensemble product.

Conclusions

Throughout the course of this study, two different calibration approaches were explored: (1) multi-objective, and (2) many-objective. In general, the best model performances were obtained from the multi-objective calibrations. Considering all of the multi-objective calibrated models, the model with the best streamflow performance was the Ensemble followed by Mosaic, SSEBop, MOD16 500 m, Noah, ALEXI, VIC, MOD16 1 km, and finally TerraClimate. Meanwhile, when considering ETa performance, the SWAT model with the best performance again used the Ensemble followed by the MOD16 500 m, Mosaic, SSEBop, Noah, VIC, MOD16 1 km, ALEXI, and again finally TerraClimate. This shows that the Ensemble used in this study had the best fit with the SWAT model and outperformed the individual ETa products.

Meanwhile, when considering the many-objective calibration without any constraints on the objective functions' values, ETa

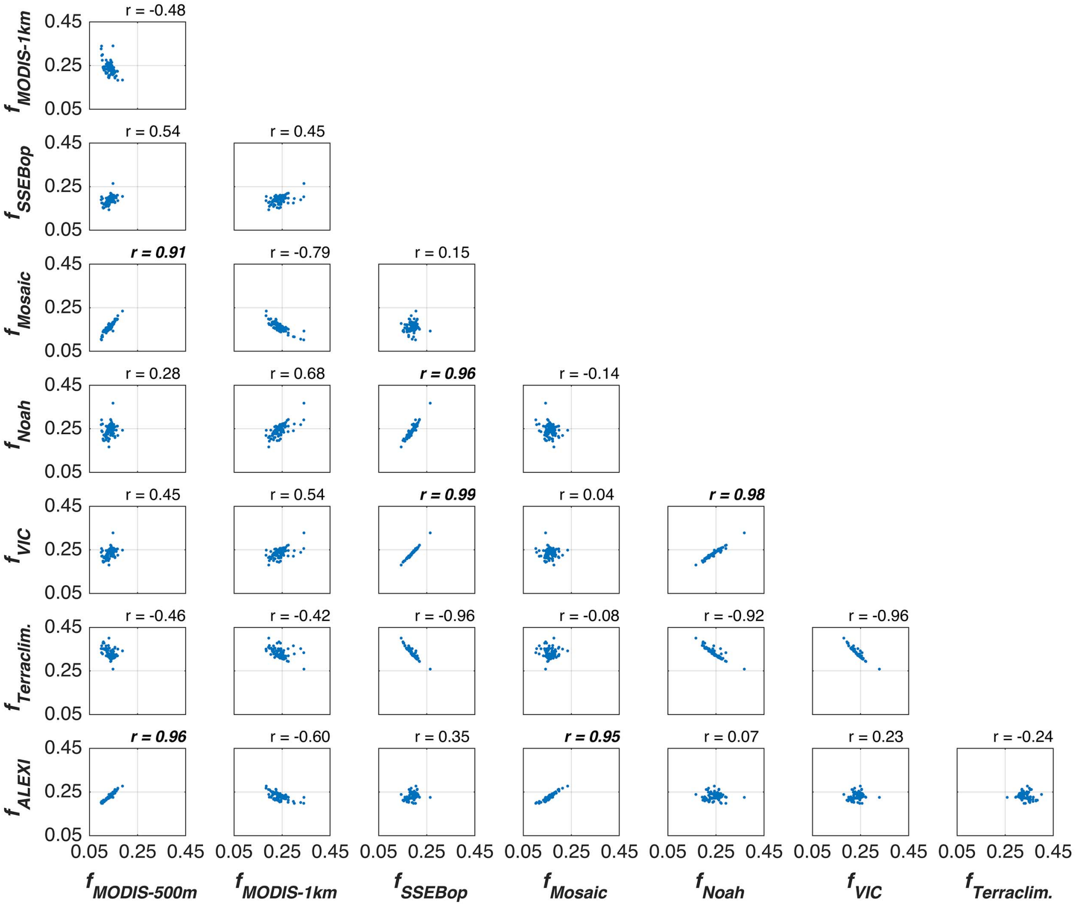

Fig. 6. Pairwise comparisons and Pearson's correlations between the ETa objective functions for the many-objective calibration using balanced weights for all nine objective functions and constrained streamflow objective function. Italic and bold numbers indicate highly correlated objective functions.

Table 4. Performance metrics for the best tradeoff solution obtained for the many-objective calibration with balanced weights and constrained streamflow objective function

Variable	Dataset	NSE	PBIAS (%)	RSR	RMSE
Flow	USGS	0.57	15.27	0.66	6.88
ETa	MOD16 500 m	0.89	9.79	0.33	13.68
	MOD16 1 km	0.76	−9.29	0.49	13.92
	SSEBop	0.83	**−27.04**	0.41	16.99
	NLDAS2-Mosaic	0.86	19.74	0.38	17.10
	NLDAS2-Noah	0.78	−18.03	0.47	16.06
	NLDAS2-VIC	0.79	**−25.40**	0.46	18.11
	TerraClimate	0.65	8.01	0.59	22.67
	USDA-ALEXI	0.79	16.73	0.45	16.08

Note: Bold values indicate unsatisfactory results.

performance was found to be satisfactory. However, the streamflow performance was not satisfactory for all Pareto-optimal solutions. This shows that the many-objective calibration is limited for SWAT model calibration when considering both streamflow and ETa simulation because the search space is much larger than the multi-objective approach. However, we found that balancing the weights in the objective functions and defining a constraint for the streamflow objective function significantly improves the calibration results for streamflow. Here we argue that the results of the many-objective calibration are more reliable than the multi-objective calibration because a calibrated model based on the multi-objective approach is biased toward one ETa product that may or may not be the most dependable for the region of study. Meanwhile, a calibrated model based on the many-objective approach tries to balance the importance of all ETa products; therefore, the outputs of the model are more reliable. In addition, correlations among the ETa objective functions show that a smaller set of ETa products should be explored in future studies, namely the MOD16 1 km, MOD16 500 m, SSEBop, and TerraClimate ETa products. The many-objective calibration also highlighted the importance of both the magnitude and the seasonal pattern in model calibration.

This study was performed for only one watershed in Michigan. Therefore, future studies should expand this work to regions with different physiographical and climatological zones. This would serve to test the robustness of the techniques implemented in this study. This would help to improve our understanding of how each ETa product performs in hydrological model calibration. Aligned with this is that only the SWAT model was used in this research, and different ETa products may fit better with different hydrological models; therefore, future studies should also explore the use of other widely used hydrological models. Furthermore, other ensemble techniques should be performed to identify the best approach for different regions.

Data Availability Statement

All data generated or used during the study are available from the corresponding author by request. This includes the generated data by SWAT models and the ones used to develop the SWAT model.

Acknowledgments

The authors thank Dr. Martha C. Anderson from USDA-ARS Hydrology and Remote Sensing Laboratory in Beltsville, MD, and Dr. Christopher R. Hain from NASA Marshall Space Flight Center at Huntsville, AL, for their help in providing the ALEXI data. This work is supported by the USDA National Institute of Food and Agriculture, Hatch project MICL02359.

Supplemental Data

Tables S1–S2 and Figs. S1–S25 are available online in the ASCE Library (www.ascelibrary.org).

References

Abatzoglou, J. T., S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch. 2018. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015." *Sci. Data* 5 (Jan). 170191. https://doi.org/10.1038/sdata.2017.191.

AmeriFlux. 2018. "AmeriFlux." Accessed September 22, 2018. http://ameriflux.lbl.gov/.

Anderson, M. C., J. M. Norman, J. R. Mecikalski, J. A. Otkin, and W. P. Kustas. 2007. "A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation." *J. Geophys. Res. Atmos.* 112 (D10): 1–17. https://doi.org/10.1029/2006JD007506.

Arnold, J. G., et al. 2012. "SWAT: Model use, calibration, and validation." *Trans. ASABE* 55 (4): 1491–1508. https://doi.org/10.13031/2013.42256.

Cammalleri, C., M. C. Anderson, F. Gao, C. R. Hain, and W. P. Kustas. 2014. "Mapping daily evapotranspiration at field scales over rainfed and irrigated agricultural areas using remote sensing data fusion." *Agric. For. Meteorol.* 186 (Mar): 1–11. https://doi.org/10.1016/j.agrformet.2013.11.001.

Climatology Lab. 2018. "Terraclimate." Accessed August 15, 2019. http://www.climatologylab.org/terraclimate.html.

Crow, W. T., E. F. Wood, and M. Pan. 2003. "Multiobjective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface radiometric temperature retrievals." *J. Geophys. Res. Atmos.* 108 (D23): 1–12. https://doi.org/10.1029/2002JD003292.

Das, I., and J. E. Dennis. 1998. "Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems." *SIAM J. Optim.* 8 (3): 631–657. https://doi.org/10.1137/S1052623496307510.

Deb, K., and H. Jain. 2014. "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints." *IEEE Trans. Evol. Comput.* 18 (4): 577–601. https://doi.org/10.1109/TEVC.2013.2281535.

Dietterich, T. G. 2000. "Ensemble methods in machine learning." In *Proc., Int. Workshop on Multiple Classifier Systems*. Berlin: Springer.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian. 2007. "Multi-model ensemble hydrologic prediction using Bayesian model averaging." *Adv. Water Resour.* 30 (5): 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014.

Einheuser, M. D., A. P. Nejadhashemi, and S. A. Woznicki. 2013. "Simulating stream health sensitivity to landscape changes due to bioenergy crops expansion." *Biomass Bioenergy* 58 (Nov): 198–209. https://doi.org/10.1016/j.biombioe.2013.08.025.

Gassman, P. W., M. R. Reyes, C. H. Green, and J. G. Arnold. 2007. "The soil and water assessment tool: Historical development, applications, and future research directions." *Trans. ASABE* 50 (4): 1211–1250. https://doi.org/10.13031/2013.23637.

Gleick, P. H. 1993. *Water in crisis*, 473. New York: Oxford University Press.

Google. 2018. "EEFlux." Accessed June 11, 2019. https://eeflux-level1.appspot.com/.

Graham, S. 1999. "Remote sensing: Feature articles." Accessed June 11, 2019. https://earthobservatory.nasa.gov/Features/RemoteSensing/.

Gupta, H. V., S. Sorooshian, and P. O. Yapo. 1999. "Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration." *J. Hydrol. Eng.* 4 (2): 135–143. https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135).

Haddeland, I., et al. 2014. "Global water resources affected by human interventions and climate change." *Proc. Natl. Acad. Sci.* 111 (9): 3251–3256. https://doi.org/10.1073/pnas.1222475110.

Hain, C. R., W. T. Crow, M. C. Anderson, and M. T. Yilmaz. 2015. "Diagnosing neglected soil moisture source–sink processes via a thermal infrared–based two-source energy balance model." *J. Hydrometeorol.* 16 (3): 1070–1086. https://doi.org/10.1175/JHM-D-14-0017.1.

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

Kim, J., B. P. Mohanty, and Y. Shin. 2015. "Effective soil moisture estimate and its uncertainty using multimodel simulation based on Bayesian model averaging." *J. Geophys. Res. Atmos.* 120 (16): 8023–8042. https://doi.org/10.1002/2014JD022905.

Kusre, B. C., D. C. Baruah, P. K. Bordoloi, and S. C. Patra. 2010. "Assessment of hydropower potential using GIS and hydrological modeling technique in Kopili River basin in Assam (India)." *Appl. Energy* 87 (1): 298–309. https://doi.org/10.1016/j.apenergy.2009.07.019.

Lee, H. S., Y. Liu, J. Ward, J. Brown, A. Maestre, H. Herr, M. A. Fresch, E. Wells, S. M. Reed, and E. Jones. 2017. "Nationwide validation of ensemble streamflow forecasts from the hydrologic ensemble forecast service (HEFS) of the US national weather service." In *Proc., AGU Fall Meeting Abstracts*. Washington, DC: American Geophysical Union.

Legates, D. R., and G. McCabe. 1999. "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation." *Water Resour. Res.* 35 (1): 233–241. https://doi.org/10.1029/1998WR900018.

Long, D., L. Longuevergne, and B. R. Scanlon. 2014. "Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites." *Water Resour. Res.* 50 (2): 1131–1151. https://doi.org/10.1002/2013WR014581.

Love, B., and A. P. Nejadhashemi. 2011. "Environmental impact analysis of biofuel crops expansion in the Saginaw River watershed." *J. Biobased Mater. Bioenergy* 5 (1): 30–54. https://doi.org/10.1166/jbmb.2011.1119.

Ma, Y., Y. Hong, Y. Chen, Y. Yang, G. Tang, Y. Yao, D. Long, C. Li, Z. Han, and R. Liu. 2018. "Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan Plateau." *J. Geophys. Res. Atmos.* 123 (2): 814–834. https://doi.org/10.1002/2017JD026648.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. 2007. "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations." *Trans. ASABE* 50 (3): 885–900. https://doi.org/10.13031/2013.23153.

Mu, Q., M. Zhao, and S. W. Running. 2011. "Improvements to a MODIS global terrestrial evapotranspiration algorithm." *Remote Sens. Environ.* 115 (8): 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019.

NASA (National Aeronautics and Space Administration). 2014. "MOD16A2 v006." Accessed June 11, 2019. https://lpdaac.usgs.gov/products/mod16a2v006/.

NASA (National Aeronautics and Space Administration). 2018a. "MOD 16—Evapotranspiration." Accessed June 11, 2019. https://modis.gsfc.nasa.gov/data/dataprod/dataproducts.php?MOD_NUMBER=16.

NASA (National Aeronautics and Space Administration). 2018b. "MOD16A2—MODIS/Terra evapotranspiration 8-day level-4 global

500 m SIN grid." Accessed June 11, 2019. https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/evapotranspiration/MOD16A2/.

NASA (National Aeronautics and Space Administration). 2018c. "NLDAS concept/goals." Accessed June 11, 2019. https://ldas.gsfc.nasa.gov/nldas/NLDASgoals.php.

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

NASS (United States Department of Agriculture National Agricultural Statistics Service). 2012. "CropScape—Cropland data layer." Accessed June 11, 2019. https://nassgeodata.gmu.edu/CropScape/.

NCDC (National Oceanic and Atmospheric Administration National Climatic Data Center). 2018. "Data tools: Local climatological data (LCD)." Accessed June 11, 2019. https://www.ncdc.noaa.gov/cdo-web/datatools/lcd.

NED (National Elevation Dataset). 2014. "The national map: Elevation." Accessed June 11, 2019. https://nationalmap.gov/elevation.html.

Neitsch, S. L., J. G. Arnold, J. R. Kiniry, and J. R. Williams. 2011. *Soil and water assessment tool theoretical documentation, version 2009*. Temple, TX: USDA-ARS Grassland, Soil and Water Research Laboratory.

NRCS (Natural Resources Conservation Service). 2014. "Web soil survey." Accessed April 26, 2017. https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx.

NTSG (University of Montana Numerical Terradynamic Simulation Group). 2018. "MODIS global evapotranspiration project (MOD16)." Accessed June 11, 2019. http://www.ntsg.umt.edu/project/modis/mod16.php.

Pan, S., H. Tian, S. R. Dangal, Q. Yang, J. Yang, C. Lu, B. Tao, W. Ren, and Z. Ouyang. 2015. "Responses of global terrestrial evapotranspiration to climate change and increasing atmospheric CO_2 in the 21st century." *Earth's Future* 3 (1): 15–35. https://doi.org/10.1002/2014EF000263.

Rajib, A., G. R. Evenson, H. E. Golden, and C. R. Lane. 2018. "Hydrologic model predictability improves with spatially explicit calibration using remotely sensed evapotranspiration and biophysical parameters." *J. Hydrol.* 567 (Dec): 668–683. https://doi.org/10.1016/j.jhydrol.2018.10.024.

Rajib, M. A., V. Merwade, and Z. Yu. 2016. "Multi-objective calibration of a hydrologic model using spatially distributed remotely sensed/in-situ soil moisture." *J. Hydrol.* 536 (May): 192–207. https://doi.org/10.1016/j.jhydrol.2016.02.037.

Rientjes, T. H. M., L. P. Muthuwatta, M. G. Bos, M. J. Booij, and H. A. Bhatti. 2013. "Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration." *J. Hydrol.* 505 (Nov): 276–290. https://doi.org/10.1016/j.jhydrol.2013.10.006.

Running, S. V., Q. Mu, M. Zhao, and A. Moreno. 2018. "User's guide MODIS global terrestrial evapotranspiration (ET) product (MOD16A2/A3 and year-end gap-filled MOD16A2GF/A3GF) NASA earth observing system MODIS land algorithm (for collection 6)." Accessed August 10, 2019. https://modis-land.gsfc.nasa.gov/pdf/MOD16UsersGuideV2.022019.pdf.

Seada, H., and K. Deb. 2016. "A unified evolutionary optimization procedure for single, multiple, and many objectives." *IEEE Trans. Evol. Comput.* 20 (3): 358–369. https://doi.org/10.1109/TEVC.2015.2459718.

Senay, G. B., S. Bohms, R. K. Singh, P. H. Gowda, N. M. Velpuri, H. Alemu, and J. P. Verdin. 2013. "Operational evapotranspiration mapping using remote sensing and weather datasets: A new parameterization for the SSEB approach." *J. Am. Water Resour. Assoc.* 49 (3): 577–591. https://doi.org/10.1111/jawr.12057.

Sharpley, A. N., and J. R. Williams. 1990. *EPIC, erosion/productivity impact calculator.* Washington, DC: USDA.

Singh, J., H. V. Knapp, J. G. Arnold, and M. Demissie. 2005. "Hydrological modeling of the Iroquois river watershed using HSPF and SWAT." *J. Am. Water Resour. Assoc.* 41 (2): 343–360. https://doi.org/10.1111/j.1752-1688.2005.tb03740.x.

Singh, R. K., and G. B. Senay. 2016. "Comparison of four different energy balance models for estimating evapotranspiration in the Midwestern United States." *Water* 8 (1): 1–19. https://doi.org/10.3390/w8010009.

Srinivasan, V., M. Konar, and M. Sivapalan. 2017. "A dynamic framework for water security." *Water Secur.* 1 (Jul): 12–20. https://doi.org/10.1016/j.wasec.2017.03.001.

Stefan, V. G., O. Merlin, S. Er-Raki, M. J. Escorihuela, and S. Khabba. 2015. "Consistency between In Situ, model-derived and high-resolution-image-based soil temperature endmembers: Towards a robust data-based model for multi-resolution monitoring of crop evapotranspiration." *Remote Sens.* 7 (8): 10444–10479. https://doi.org/10.3390/rs70810444.

Tebaldi, C., and R. Knutti. 2007. "The use of the multi-model ensemble in probabilistic climate projections." *Philos. Trans. A. Math. Phys. Eng. Sci.* 365 (1857): 2053–2075. https://doi.org/10.1098/rsta.2007.2076.

Texas A&M University. 2017. "SWAT: Soil and water assessment tool." Accessed June 11, 2019. http://swat.tamu.edu/.

Tian, D., and H. Medina. 2017. "Comparative assessment of several post-processing methods for correcting evapotranspiration forecasts derived from TIGGE datasets." In *Proc., AGU Fall Meeting Abstracts*. Washington, DC: American Geophysical Union.

Tobin, K. J., and M. E. Bennett. 2017. "Constraining SWAT calibration with remotely sensed evapotranspiration data." *J. Am. Water Resour. Assoc.* 53 (3): 593–604. https://doi.org/10.1111/1752-1688.12516.

USGS. 2016a. "Evapotranspiration—The water cycle." Accessed June 11, 2019. https://water.usgs.gov/edu/watercycleevapotranspiration.html.

USGS. 2016b. "USGS current conditions for USGS 04155500 Pine river near Midland, MI." Accessed June 11, 2019. https://waterdata.usgs.gov/mi/nwis/dv?referred_module=sw&site_no=04155500.

USGS. 2018. "National water information system: Web interface." Accessed June 11, 2019. https://waterdata.usgs.gov/usa/nwis/sw.

Velpuri, N. M., G. B. Senay, R. K. Singh, S. Bohms, and J. P. Verdin. 2013. "A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET." *Remote Sens. Environ.* 139 (Dec): 35–49. https://doi.org/10.1016/j.rse.2013.07.013.

Vervoort, R. W., S. F. Miechels, F. F. van Ogtrop, and J. H. Guillaume. 2014. "Remotely sensed evapotranspiration to calibrate a lumped conceptual model: Pitfalls and opportunities." *J. Hydrol.* 519 (Nov): 3223–3236. https://doi.org/10.1016/j.jhydrol.2014.10.034.

Wallis, T. W., and J. F. Griffiths. 1995. "An assessment of the weather generator (WXGEN) used in the erosion/productivity impact calculator (EPIC)." *Agric. For. Meteorol.* 73 (1–2): 115–133. https://doi.org/10.1016/0168-1923(94)02172-G.

Wambura, F. J., O. Dietrich, and G. Lischeid. 2018. "Improving a distributed hydrological model using evapotranspiration-related boundary conditions as additional constraints in a data-scarce river basin." *Hydrol. Processes* 32 (6): 759–775. https://doi.org/10.1002/hyp.11453.

Wanders, N., M. F. Bierkens, S. M. de Jong, A. de Roo, and D. Karssenberg. 2014. "The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models." *Water Resour. Res.* 50 (8): 6874–6891. https://doi.org/10.1002/2013WR014639.

Wang, S., B. C. Ancell, G. H. Huang, and B. W. Baetz. 2018. "Improving robustness of hydrologic ensemble predictions through probabilistic pre-and post-processing in sequential data assimilation." *Water Resour. Res.* 54 (3): 2129–2151. https://doi.org/10.1002/2018WR022546.

Woznicki, S. A., and A. P. Nejadhashemi. 2012. "Sensitivity analysis of best management practices under climate change scenarios." *J. Am. Water Resour. Assoc.* 48 (1): 90–112. https://doi.org/10.1111/j.1752-1688.2011.00598.x.

Xia, Y., M. T. Hobbins, Q. Mu, and M. B. Ek. 2015. "Evaluation of NLDAS-2 evapotranspiration against tower flux site observations." *Hydrol. Processes* 29 (7): 1757–1771. https://doi.org/10.1002/hyp.10299.

Yao, Y., et al. 2017. "Improving global terrestrial evapotranspiration estimation using support vector machine by integrating three process-based algorithms." *Agric. For. Meteorol.* 242 (Aug): 55–74. https://doi.org/10.1016/j.agrformet.2017.04.011.

Zeleny, M., and J. L. Cochrane. 1973. *Multiple criteria decision making*. Columbia, SC: University of South Carolina Press.

Zhang, Y., F. H. S. Chiew, L. Zhang, and H. Li. 2009. "Use of remotely sensed actual evapotranspiration to improve rainfall–runoff modeling in Southeast Australia." *J. Hydrometeorol.* 10 (4): 969–980. https://doi.org/10.1175/2009JHM1061.1.

Zhao, Y., L. Zhixiang, and W. Yongping. 2019. "An assessment of global precipitation and evapotranspiration products for regional applications." *Remote Sens.* 11 (9): 1077. https://doi.org/10.3390/rs11091077.