JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 56, No. 4 AMERICAN WATER RESOURCES ASSOCIATION August 2020

Analyzing the Variability of Remote Sensing and Hydrologic Model Evapotranspiration Products in a Watershed in Michigan

Matthew R. Herman, A. Pouyan Nejadhashemi, Juan Sebastian Hernandez-Suarez, and Ali M. Sadeghi

Research Impact Statement: This study can help policy makers and researchers select the best actual evapotranspiration dataset for different tasks such as monitoring of agricultural lands or tracking deforestation.

ABSTRACT: The goal of this paper was to statistically explore the spatiotemporal performance of remotely sensed actual evapotranspiration (ETa) datasets and a remotely sensed ensemble in a region that lacks observed data. The remotely sensed datasets were further compared with ETa results from a physically based hydrologic model (Soil and Water Assessment Tool) to examine the differences and determine the level of agreement between the ETa datasets and the model outputs. ETa datasets were compared on temporal (i.e., monthly and seasonal basis) and spatial (i.e., landuse) scales at both watershed and subbasin levels. The results showed a lack of consistent similarities and differences among the datasets when evaluating the monthly ETa variations; however, the seasonal aggregated data presented more consistent similarities and differences during the spring and summer compared to the fall and winter. Meanwhile, spatial analysis of the datasets showed the MOD16A2 500 m ETa product was the most versatile of the tested datasets, being able to differentiate between landuses during all seasons. Furthermore, the use of an averaging ensemble was able to improve overall ETa performance in the study area. This study showed that the remotely sensed ETa products are not similar throughout the year, but the appropriate application periods for different ETa products were identified. Finally, spatial variabilities of the ETa products are more in tune with landuse and climate characteristics.

(KEYWORDS: evapotranspiration; remote sensing; average ensemble; hydrological model; spatial variability; temporal variability.)

INTRODUCTION

Freshwater is vital for life and therefore understanding how the hydrological cycle changes has become a major focus of many researchers, especially given the increased demand for water across the globe (Clark et al. 2015; Srinivasan et al. 2017). Traditionally, this has been accomplished using monitoring stations that record different aspects of the hydrological cycle, such as streamflow and precipitation. However, these stations can be expensive to install, maintain, and operate and thus their coverage is often not adequate to capture spatiotemporal variabilities of hydrological cycle elements especially over large areas (Wanders et al. 2014). One solution to this is the use of remote sensing (RS) products. RS is the use of sensors and tools to indirectly measure the characteristics of an object (Graham 1999). In addition, with the advancement of satellite technology, generating consistent global monitoring datasets such as different elements of the hydrological cycle in forms of remotely sensed products has become common (Long et al. 2014).

Paper No. JAWRA-19-0014-P of the *Journal of the American Water Resources Association* (JAWRA). Received March 25, 2019; accepted April 5, 2020. **Discussions are open until six months from issue publication**.

Department of Biosystems and Agricultural Engineering (Herman, Nejadhashemi, Hernandez-Suarez), Michigan State University, East Lansing, Michigan, USA; and Hydrology and Remote Sensing Laboratory (Sadeghi), USDA Agricultural Research Service, Beltsville, Maryland, USA (Correspondence to Nejadhashemi: pouyan@msu.edu).

Citation: Herman, M.R., A.P., Nejadhashemi, J.S., Hernandez-Suarez, and A.M., Sadeghi. 2020. "Analyzing the Variability of Remote Sensing and Hydrologic Model Evapotranspiration Products in a Watershed in Michigan." *Journal of the American Water Resources Association* 56 (4): 738–755. https://doi.org/10.1111/1752-1688.12849.

In the hydrological cycle, evapotranspiration (ET) is an influential component since it is the measure of how much water enters the atmosphere from both the Earth's surface and from plants (USGS 2016a). Which means that ET supplies water vapor to the atmosphere driving weather patterns and precipitation distributions (Pan et al. 2015; USGS 2016b). Meanwhile, since ET measures the loss of moisture from plants and soil, its magnitude is dependent on the landscape. Therefore, measuring ET on a large scale is difficult through traditional techniques (Wu et al. 2008), but is a prime hydrological component to be measured through remotely sensed techniques (Anderson et al. 2012). This has led to the development of a variety of different ET remotely sensed monitoring products, such as the Simplified Surface Energy Balance (SSEB) (Zhang et al. 2016), the Atmosphere-Land Exchange Inverse (ALEXI) (Anderson et al. 2007; Senay et al. 2013), the Moderate Resolution Imaging Spectroradiometer (MODIS) Global Evapotranspiration Project (MOD16) (Zhang et al. 2016; NTSG 2018), the Google Earth Engine Evapotranspiration Flux (Google 2018), and the North American Land Data Assimilation Systems phase 2 (NLDAS-2) (Xia et al. 2015). These products can be categorized based on the method they use to calculate ET with the most common categories being Surface Energy Balance Methods, Penman–Monteith Methods, and Priestley–Taylor Methods (Bhattarai et al. 2016; Zhang et al. 2016). However, each of these methods have different assumptions and required inputs to calculate ET as well as a higher level of uncertainty associated with the remotely sensed data compared to traditional ET monitoring techniques (van der Tol and Parodi 2012; Zhang et al. 2016). All of these can make it challenging for researchers and policy makers to know which ET product should be used considering landuse/landcover and a period of study. To address this issue several papers have been published that have compared different ETa techniques and datasets (Mueller et al. 2011; Vinukollu et al. 2011; Ershadi et al. 2015; Singh and Senay 2016; Zhang et al. 2016). Mueller et al. (2011) evaluated 41 different ET products at the global scale. The datasets they compared were divided into four categories of diagnostic datasets, Land Surface Models, Reanalyses, and IPCC AR4 Simulations. The results of this analysis showed that many of the tested datasets share global spatial patterns of high and low ET; however, the IPCC AR4 simulations fluxed more than other datasets across the globe (Mueller et al. 2011). Vinukollu et al. (2011) evaluated three ET products at the global scale, namely a Surface Energy Balance System (SEBS), a Penman–Monteith algorithm, and a Priestley–Taylor algorithm by comparing them to data from 12 eddy covariance towers. The results of this analysis showed that the Priestley–Taylor algorithm had the highest correlations to tower observations among the ET products followed by the SEBS and then the Penman–Monteith algorithm (Vinukollu et al. 2011). Singh and Senay (2016) compared ETa products from four different models, namely the Mapping EvapoTranspiration at high Resolution with Internalized Calibration (METRIC) model, Surface Energy Balance Algorithm for Land model, SEBS model, and the Operational SSEB (SSEBop) model. The results showed all models performed well at the observed data sites. In addition, it was noted that increased model complexity does not necessarily mean better performance as the most complex METRIC model and the simpler SSEBop model had similar performances (Singh and Senay 2016). This finding was also found in Ershadi et al. (2014), which compared the performance of four commonly used ETa land surface models, namely the SEBS, Penman–Monteith, Advection–Aridity, and Modified Priestley–Taylor, across multiple landuses through the use of FLUXNET towers. The results showed that the Modified Priestley–Taylor approach had the best fit with the FLUXNET towers followed by SEBS, Penman–Monteith, and finally Advection–Aridity (Ershadi et al. 2014). Ershadi et al. (2014) also identified the performance of each of the four ETa approaches with respect to the landuse surrounding each FLUXNET tower. This highlighted the spatial nature of ETa and showed that it is not possible to select a single ETa technique that perfoms well for all landuses. Ershadi et al. (2015) expanded on this result by testing how Penman–Monteith ETa models responded to changes in model structure and parameterization. Again, ETa models were compared to FLUXNET towers, and even among models based on the same technique (Penman–Monteith) no single model had superior performance across all biomes selected (Ershadi et al. 2015). Xia et al. (2015) compared the ETa products from the four NLDAS-2 land surface models (Mosaic, Noah, Sacramento Soil Moisture Accounting Model [SAC-SMA], and Variable Infiltration Capacity [VIC]) at 29 AmeriFlux sites and 14 Atmospheric Radiation Measurement/Cloud and Radiation Testbed sites. The results showed that all four models were able to capture broad scale ETa features. However, among the models there was considerable variability, with the Mosaic model overestimating ETa values, Noah ETa values were the closest to the observed data, and the SAC-SMA and VIC ETa values were between the Mosaic and Noah (Xia et al. 2015). While these studies compare ETa datasets to observed data, they are often limited to specific points and given the spatial variability in ETa across the landscape these results do not necessarily indicate which ETa product performs best in all locations. Meanwhile, Zhang et al. (2016) explored the methodologies used to calculated ETa, including Surface Energy Balances, Penman–Monteith Methods, Priestley–Taylor Methods, T_s-VI Space Methods, the MEP Method, Water-Carbon Linkage Methods, Water Balance Methods, and Empirical Models. Throughout their research of these different techniques, different advantages and limitations were identified, ranging from ease of use and data requirement to the application of unique linkages and the impact of cloud cover. Ultimately, no single best technique was identified; however, a lack of studies focused on identifying the sources of uncertainty for ETa calculation methodologies was noted.

One technique to address the uncertainty within remotely sensed datasets is the use of an ensemble of several different products (Duan et al. 2007). Creating an ensemble of datasets helps reduce the uncertainty of individual datasets by combining the benefits of each dataset while minimizing negative aspects like outliers (Dietterich 2000). This has led to the creation of a variety of ensemble techniques and applications that have been applied to remotely sensed products (Christensen and Lettenmaier 2006; Fowler and Ekström 2009; Lee et al. 2017; Wang et al. 2018). The complexity of these techniques ranges from very simple calculations such as simple averaging to very complex techniques such as ensemble Kalman filter (EnKF) (Giorgi and Mearns 2003; Ershadi et al. 2014; Kim et al. 2015; Wang et al. 2018). However, the Bayesian Model Averaging (BMA) is the most commonly used ensembling technique for ET remotely sensed products (Kim et al. 2015; Tian and Medina 2017; Yao et al. 2017; Ma et al. 2018) that reduces overall dataset uncertainty by weighting ET products based on the observed data (Kim et al. 2015). However, this technique is dependent on the availability of observed data, which depending on the region can be difficult to obtain.

In summary, the wide range of techniques can make it challenging to know which technique should be applied. Therefore, given the challenges associated with the selection and use of remotely sensed ET products in the field of hydrology three objectives were identified for this study: (1) explore the temporal variability of individual and average ensemble remotely sensed datasets by comparing monthly mean differences at both watershed and landuse scales; (2) evaluate the spatial variability of individual and average ensemble remotely sensed datasets by comparing mean differences among landuses for each individual ETa dataset and across all ETa datasets; (3) compare the mean differences in individual remotely sensed datasets to the average ensemble and hydrological model's outputs at both watershed and subbasin scales.

MATERIALS AND METHODS

To accomplish the objectives of this study a variety of tasks were performed. First eight remotely sensed ETa datasets along with an average ETa ensemble and ETa output of a hydrological model were obtained for a study area (Table 1). Since each of these datasets has different spatial and temporal resolutions, they were aggregated or disaggregated to create a series of comparable ETa datasets. To determine their performance in the study area, several forms of statistical analysis were performed to examine the temporal spatial variabilities in addition to their fit to the average ensemble and hydrological model output. The following sections provide additional information about all of the processes used in this study.

Study Area

The Honeyoey Creek-Pine Creek Watershed (Hydrologic Unit Code 0408020203) with the area of 1,100 km^2 was selected for this study (Figure 1). Located in Michigan's Lower Peninsula, this watershed is part of the Saginaw Bay Watershed, which is the largest watershed in Michigan with final outlet at Lake Huron. Furthermore, this region has been identified as an area of concern by the United States (U.S.) Environmental Protection Agency due to the degradation of fisheries, presence of contaminated sediments, and implementation of fish consumption advisories within the region (EPA 2017). On average, the region receives 81 cm of precipitation per year with higher precipitations observed during the months between April and November (U.S. Climate Data 2018). In addition to seasonal variation in precipitation, there is also seasonal variation in the type of storm systems that deliver precipitation to the area. Throughout the wetter months of the year (April through November) convection storm fronts driven by higher surface temperatures and high humidity are more common. Meanwhile, during the dryer months (December through March) mid-latitude cyclones and lake-effect snow belts are responsible for most of the precipitation. Furthermore, the late fall and winter months (November through February) experience more clouds and shorter days, whereas the late spring and summer months experience fewer clouds and longer days. Meanwhile, air temperature in the region ranges from −10 to 27 , with the winter months (December through February) having colder temperatures and snow, whereas the summer months (June through August) have

TABLE 1. Summary of remotely sensed actual evapotranspiration (ETa) datasets used in this study.

ETa dataset	Coverage	Resolution		Accuracy (mm/month)	Reference
		Spatial (km^2)	Temporal		
SSEBop	Contiguous United States	1.0	Monthly	27.90	Velpuri et al. (2013)
ALEXI	Contiguous United States	4.0	Daily	30.15	Cammalleri et al. (2014)
MOD16A2 1 km	Global	0.5	Eight-day	1.25	Mu et al. (2011)
MOD16A2 500 m	Global	1.0	Eight-day	1.25	(Mu et al. (2011)
NLDAS-2: Mosaic	North America	12.0	Hourly/monthly	6.00	Long et al. (2014)
NLDAS-2: Noah	North America	12.0	Hourly/monthly	3.66	(Long et al. (2014)
NLDAS-2: VIC	North America	12.0	Hourly/monthly	6.66	Long et al. (2014)
TerraClimate	Global	4.0	Monthly	4.75	Abatzoglou et al. (2018)

Notes: SSEBop, Operational Simplified Surface Energy Balance; ALEXI, Atmosphere-Land Exchange Inverse; MOD16A2, MODIS Global Evapotranspiration Project; NLDAS-2, North American Land Data Assimilation Systems phase 2; VIC, Variable Infiltration Capacity.

hotter temperatures and more rainfall (U.S. Climate Data 2018). In general, the relative humidity for the region ranges from the upper sixty to mid seventy percentages, however the high temperatures during the summer months mean that the atmosphere is holding more water resulting in a more humid period, whereasthe winter months are less humid. Soils in the area are dominated by mixtures of loam and sand with low slopes (NRCS 2018). Landuse in the Honeyoey watershed is dominated by agricultural land (~57%) followed by forests (~23%), wetlands (~17%), and urban areas (~3%). Given the heavy agricultural nature of the region is it important to note that corn and soybean rotations are the most common crops; however, eight different cropping systems have been identified in the region including alfalfa, corn, field peas, hay, pasture, sugar beet, soybean, and winter wheat. In general, agricultural operations like tillage and crop planning start in the mid spring (i.e., May) and crops are harvested in the mid fall (i.e., October) (Love and Nejadhashemi 2011). In cases where cover crops are utilized, planting begins postharvest in the fall, which requires additional tillage and planting operations (Love and Nejadhashemi 2011). Overall, 13 types of landuses were identified (NASS 2018). These individual landuses were also combined into four major landuse. Figure 2 shows the spatial distribution of the major landuse categories throughout the Honeyoey watershed.

Meanwhile regarding hydrological and climatological monitoring in the area, streamflow is monitored by a U.S. Geological Survey (USGS) station located at the outlet of the region (Figure 1). Furthermore, two precipitation and two temperature National Climatic Data Center stations are located within the Honeyoey watershed (NCDC 2018) (Figure 1). Automated airport weather stations are also located within and around the Honeyoey watershed and collect wind speed and direction, temperature, dew point, altimeter setting, density altitude, visibility, sky condition, cloud ceiling, precipitation, and precipitation type (FAA 2018). Additional weather stations from the MSU Enviroweather system measure air and soil temperature, precipitation, relative humidity, wind speed and direction, solar radiation, leaf wetness, and potential ET (Enviroweather 2018). However, none of the enviroweather stations are located within the study region. Meanwhile, there are several AmeriFlux stations located in Michigan that can be used to report ETa; however, the closest of these stations is 116 km from the Honeyoey watershed (AmeriFlux 2018).

RS ET Products

To examine the spatial and temporal performance of remotely sensed ET products, eight actual ET (ETa) datasets were obtained for the study area. ETa describes the actual amount of water loss that occurs at a site via evaporation and transpiration and thus is limited by the actual amount of water and energy present (McVicar et al. 2012; NOAA 2017). The ETa datasets utilized for this study include (1) the USGS SSEBop, (2) the ALEXI, (3) the MODIS MOD16A2 500 m, (4) the MOD16A2 1 km, (5) the NLDAS-2 Evapotranspiration Mosaic, (6) the NLDAS-2 Noah, (7) the NLDAS-2 VIC, and finally (8) TerraClimate. These ETa products were selected for this study since they are widely used by the scientific community for water resources research. In addition, all of the products are available for the period of study and the area of interest. Additional information about these products are provided in Supporting Information. Table 1 summarizes the spatial and temporal resolutions for each of the ETa datasets.

Meanwhile, since the ETa were obtained from remotely sensed products, calibration and validation were necessary before the products were made available to the public. This was performed for all of the

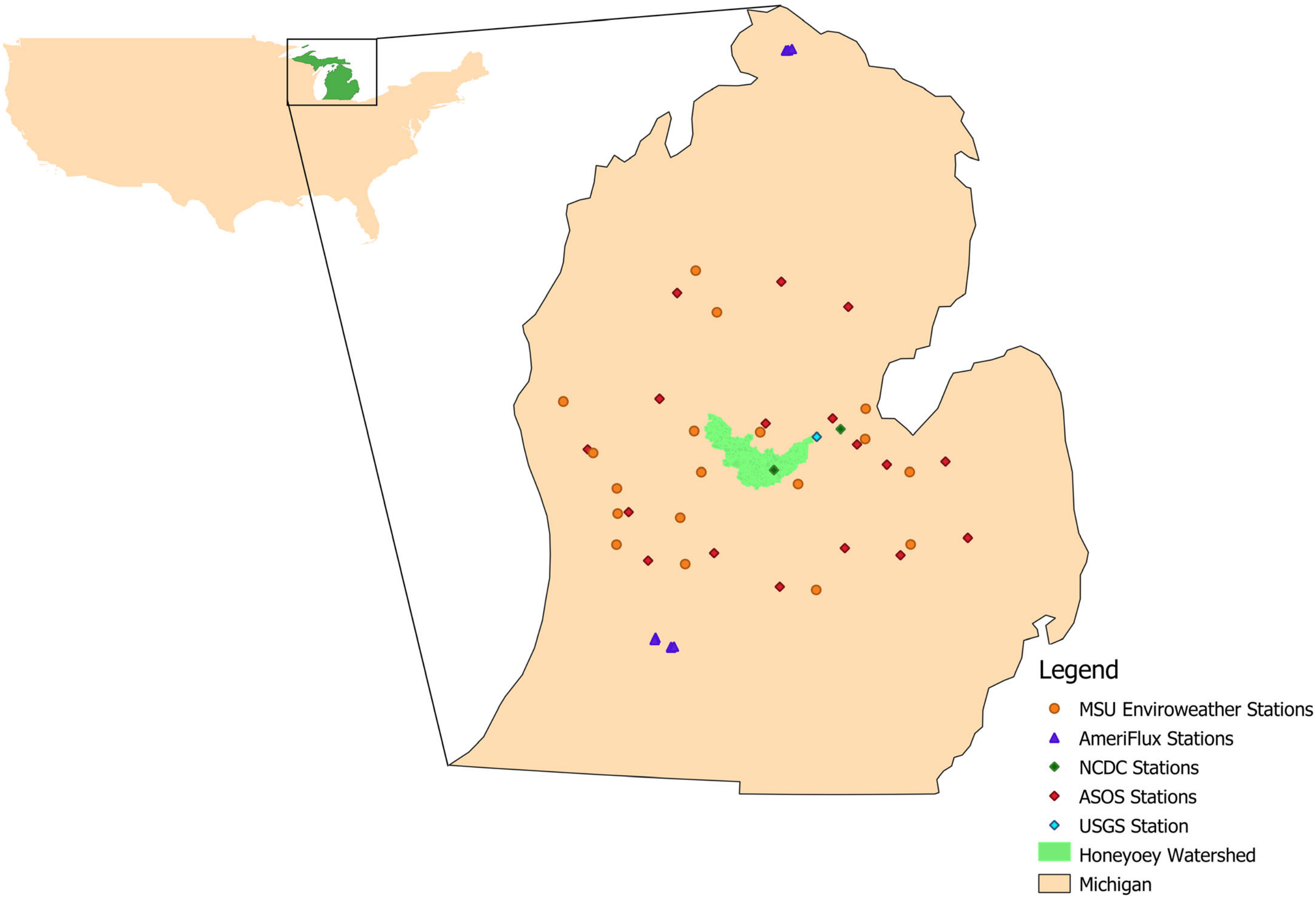

FIGURE 1. Map of the Honeyoey watershed and locations of climatological stations within and near the region.

aforementioned ETa products by the agencies supplying the data and the levels of accuracy were also reported in Table 1. As can be seen in Table 1, the ETa accuracy of the remotely sensed datasets varies between 1.25 to 30.15 mm/month and with the MOD16A2 datasets having the smallest error followed by NLDAS-2: Noah, TerraClimate, NLDAS-2: Mosaic, NLDAS-2: VIC, SSEBop, and finally ALEXI. However, it is important to note that these errors are based on site specific comparisons with observed data. This means that for any given location between two observed sites, the actual error associated with each dataset could flux. In addition, the accuracy level reported in Table 1 are not absolute errors, which mean that they can change throughout the years and for different landuses. Therefore, in case that measured ETa data are available for the study site, validation of the ETa RS products in use with the measured Eta data is highly recommended. Given this, it is important to note that the goal of this study is not to perform revalidation for the selected datasets but to see how the ETa datasets perform within the study area. However, we are interested to assess how different spatial and temporal variation are represented by each dataset while identifying the possible sources of discrepancy among datasets. In addition, and as presented in Study Area section of this paper, while there are many monitoring sites within and around the study area, there is a lack of observed ETa datasets. Therefore, to help to account for the uncertainty within the datasets, an ensemble dataset based on an averaging technique (Tebaldi and Knutti 2007) was also created. It is important to note that, the use of a straight average for ensembling is not as robust as other techniques such as BMA (Krishnamurti et al. 2000); however, due to the lack of observed data in the region (Figure 2) it was considered as the most appropriate technique to use. Therefore, in this study, the arithmetic mean method was used to calculate the average ETa ensemble values.

Hydrological Model

In addition to the eight remotely sensed ETa products and the average ETa ensemble, a hydrological

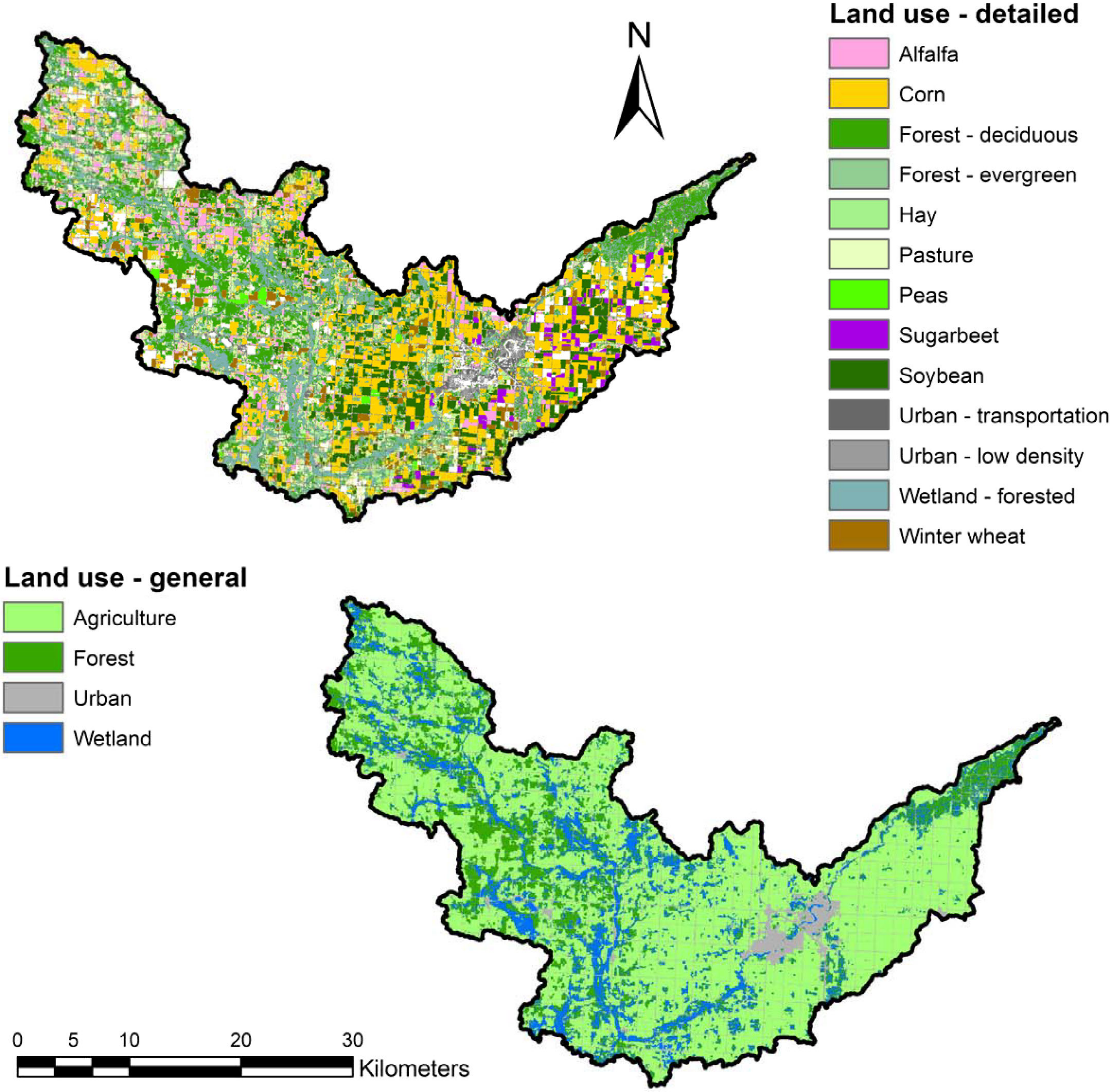

FIGURE 2. Map of the individual (a) and major (b) landuse classes within the Honeyoey watershed based on the 30 m resolution map obtained from the Cropland Data Layer developed by the United States Department of Agriculture-National Agricultural Statistics Service.

model was used to estimate ETa for the region as well. Hydrological models are often used to simulate the hydrological cycle across the landscape, since they are an efficient and inexpensive alternative to monitoring (Giri et al. 2012). They accomplish this, in general, by performing a water balance for the region, which utilizes various calculations describing water movement throughout the landscape as well as the interactions between water and biotic and abiotic characteristics (Martinez-Martinez et al. 2014). ET is one of the major components of the water balance and as such plays a major role in hydrological models. To estimate ETa, hydrological models often first calculate potential ET and then account for actual loss by determining the impacts of landcover and soil moisture (Kite and Droogers 2000).

In this study, the hydrological model selected was the Soil and Water Assessment Tool or SWAT model.

SWAT is a semi-distributed, continuous-time hydrological model developed by the U.S. Department of Agriculture (USDA) — Agriculture Research Service (ARS) and Texas A&M AgriLife Research (Texas A&M University 2018). This is the most widely used hydrologic model, which utilizes several different datasets, such as topography, soil properties, landuse, and climatological observations to simulate hydrological parameters such as streamflow and ET (Gassman et al. 2007). Regarding ETa estimation, the SWAT model first calculates potential ET. This can be done one of three ways, namely the Penman–Monteith method, the Priestley–Taylor method, and the Hargreaves method, with the Penman–Monteith Method as the default (Neitsch et al. 2011). After this, SWAT takes into account 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 calculations are performed at the hydrologic response unit scale, which divided the region into subbasins that have unique physiographical characteristics. For this study, 250 subbasins were created in the study area. This number was selected due to limitations in the number of unique landuses that could be applied within the SWAT model. Ultimately, all of these calculations result in the creation of a dataset that reports monthly ETa at the subbasin level.

To ensure that the hydrological model represented the study area, calibration and validation were performed for the period of 2003–2014, with 2003–2004 serving as a model warmup period, 2005–2009 used as the calibration period, and 2010–2014 used as the validation period. The hydrological cycle component used for this process was streamflow, with the daily model streamflow output being compared to observed daily streamflow at the watershed outlet. To evaluate this comparison, three statistical criteria were used, namely Nash–Sutcliffe efficiency (NSE), percent bias (PBIAS), and root mean squared error-observations standard deviation ratio (RSR), which were identified by Moriasi et al. (2007). Calibration and validation were successful for the developed model since the following ranges for each statistical criterion were met: NSE >0.5, PBIAS ±25, and RSR <0.7 (Herman et al. 2015).

Remotely Sensed Actual ET Data Source and Conversion Procedure

All the ETa datasets were obtained for the period 2003–2014 for the study area. This period was selected since all of the selected ETa datasets had data available during this period. The NLDAS-2 datasets (from Mosaic, Noah and VIC models) were obtained using the NASA Goddard Earth Science Data and Information Services Center website (NASA/GSFG 2018). ETa values for each model were extracted using the wgrib program developed by the NOAA's National Centers for Environmental Prediction (NOAA-NCEP 2013). Similarly, average ETa values from MOD16A (eight-day values, 0.5 and 1 km resolutions) and TerraClimate (monthly values) datasets were obtained using the code editor of Google Earth Engine (Gorelick et al. 2017). Missing eight-day values in MOD16A datasets were completed using multiyear averages for either the respective week or month of the missing values (the latter when the average week value was not available) (Mu et al. 2011). Meanwhile, averaged ETa values for the SSEBop product and the NASA ALEXI product were obtained from USGS and USDA, respectively. The SSEBop product (available once every eight days) was derived from 1 km MODIS thermal imagery, whereas the ALEXI product (available daily) was derived from the 4 km thermal images that were obtained from Geostationary Operational Environmental Satellites (Senay et al. 2013; Velpuriet al. 2013; Hain et al. 2015). Then ETa values were averaged to create monthly ETa values.

However, to compare these ETa datasets, they need to be converted to the similar spatial and temporal resolutions. The first step was to ensure that each ETa dataset was reported on a monthly basis. For datasets already reported on a monthly basis (SSEBop, NLDAS-2, and TerraClimate) no processing was needed. However, for datasets that reported ETa on a daily (ALEXI) or eight-day (MOD16) basis, values within each month were summed using the R programing language. The second step was to ensure that each ETa dataset accounted for spatial variability within the landscape. To accomplish this the ETa datasets had to be averaged for each subbasin. This was done using weighted area averaging technique on all ETa datasets for each unique physiographic region within the Honeyoey watershed. The average ETa values were obtained by resampling the raster files to a cell size of 10 m and computing the mean value of the cells within each of the 250 subbasins. Weighted area averaging was used since it was able to address the issues of multiple pixels and partial pixels occurring within unique physiographic regions and resulted in a single ETa value for each subbasin. By performing these two processes, eight monthly ETa datasets at the subbasin level were created that can be used for further analysis.

Statistical Analysis

To compare the performance of the eight datasets, average ensemble, and SWAT model within the study region, three different statistical approaches were used. These analyses were performed to take into account different spatial scales (subbasin, watershed), landuses (major and individual), and temporal resolutions (overall, seasonally, and monthly).

The first statistical approach evaluated the temporal variability of the different ETa datasets and utilized multi-pairwise comparisons to estimate the monthly mean differences between ETa datasets, for both the whole watershed and for specific landuse types. This was done to determine any consistent similarities and differences that could be found among the ETa datasets throughout the year. Since this was done for both the entire watershed and for each landuse, a total of 18 area-weighted ETa monthly time series were generated for each ETa dataset (13 individual landuses, 4 major landuses, and the entire watershed). To compare these datasets, two different models were used: (1) for overall comparisons the Generalized Least-Square (GLS) estimation with Autoregressive model, with a lag of 1, or AR (1) was used since complete time series was being compared (Fox and Monette 2002); (2) while for monthly and seasonal comparisons the GLS estimation with Continuous Autoregressive model with lag 1, or CAR (1) was used since irregularly spaced time series were being compared (Wang 2013). In both cases, the difference between two analyzed time series was used as the response variable. Furthermore, a $p < 0.05$ denoted datasets that were significantly different (Nejadhashemi et al. 2012).

The second statistical approach evaluated the spatial variability of the different ETa datasets and utilized multi-pairwise comparisons of different landuse types (including the watershed average) for each individual ETa dataset and across all ETa datasets. By performing both of these analyses, it is possible to evaluate the performance of individual ETa datasets in differentiating among landuse classes as well as determine if different ETa datasets perform in a similar manner for individual landuses. This again utilized both GLS estimation with AR (1) and GLS CAR (1) in which area-weighted ETa monthly time series are obtained for the whole watershed. Meanwhile, each landuse were pairwise compared. Again, the difference between two time series is used as the response variable.

Finally, for the third approach, we computed the mean difference between the ETa datasets with respect to the average ensemble and SWAT model for each subbasin. In this case, the GLS-AR (1) regression method was used to perform an overall comparison, which reported the mean difference and p-value for each subbasin. As a result, a series of maps were created that represented the spatial variation in the mean differences with respect to the average ensemble and SWAT model ETa values.

RESULTS AND DISCUSSION

In this section of the paper, several tables were used to identify similarities and differences among the ETa datasets. When different datasets have the same superscript, it indicates that the mean difference of the datasets is not statistically different from zero. Meanwhile, if two datasets have different superscripts, it indicates that while each dataset is similar to another dataset, the mean difference between them is statistically different from zero. Finally, if a dataset has no superscript, it indicates that the mean difference of that dataset and all other datasets is statistically different from zero. It is important to note that in the context of this text, the term "clusters" is used to describe sets of ETa datasets for which the mean difference is significantly not different from zero.

Temporal Statistical Analysis

Monthly Analysis. Watershed Scale Analysis. Temporal cluster analysis was performed to determine if any of the ETa datasets produced similar results during specific times of the year. Table 2 presents the mean monthly values of each ETa dataset for the entire Honeyoey watershed as well as any similarities between datasets with superscripted letters.

Regarding similarities between datasets throughout the course of the year, no noticeable consistent similarities and differences were seen. This is likely due to the fact that each of the ETa datasets utilize different equations, approaches, and spatial resolutions. Another possible cause for the lack of consistent similarities and differences among the ETa datasets is the fact that this analysis is the summary over the entire watershed, and consistent similarities and differences found within specific landuses could be lost due to data aggregation at the watershed level.

Landuse Analysis. To determine if consistent similarities and differences among the ETa datasets were lost due to aggregation at the watershed scale, monthly analysis was also performed for the major landuses (agriculture, forest, urban, and wetland) as well as all of the individual landuses (ALFA, CORN, FPEA, FRSD, FRSE, HAY, PAST, SGBT, SOYB, URLD, UTRN, WETF, WWHT). Tables S1–S4 showing the mean monthly values of each ETa dataset for

TABLE 2. Average monthly ETa values for each dataset for the entire watershed with clusters indicated by superscripts for each column.

	Month											
Datasets	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
MOD16A2 1 km	16.06^{a}	21.77^{a}	37.24^{a}	37.99	59.82^{a}	83.85^{a}	100.06^{a}	83.16^{a}	44.02	25.29^{a}	20.22^{a}	15.12^{a}
MOD16A2 500 m	15.81^{a}	21.14^{a}	37.55^{a}	44.98^{a}	77.94^{b}	109.21^{b}	130.74^{b,c}	110.68^{b}	57.27^{a,b}	26.58^{a}	17.14	10.72^{b}
SSEBop	0.03	0.01	10.38^{b}	26.82^{b}	50.02^{c,d}	92.37^{a,c,d}	117.93^{d}	99.96^{a,c}	52.34^{a,c}	12.16	5.77^{b}	0.71^{c}
NLDAS-2:Mosaic	10.91^{b,c}	11.86^{b}	26.84^{c,d}	58.93^{c}	95.36	119.00^{e}	135.66^{b}	115.18^{b}	85.01	49.21^{b}	21.83^{a}	11.76^{b}
NLDAS-2:Noah	10.21^{b}	12.53^{b}	19.11^{e}	28.36^{b}	43.84^{c}	74.62	102.31^{a}	99.85^{a,c}	67.05^{d}	28.46	10.55	7.36^{d}
NLDAS-2:VIC	7.61	9.77	10.19^{b}	15.40	48.11^{d}	89.21^{a,c}	116.70^{d}	97.60^{a,c}	50.54^{c}	16.37	6.00^{b}	7.14^{d}
TerraClimate	—[1]	—[1]	18.00^{b,c,d,e}	81.94	101.65	110.67^{b,e}	97.78^{a,e}	87.19^{a,c}	65.24^{a,b,c,d}	49.53^{b}	22.52^{a}	1.40^{c}
ALEXI	22.96	37.32	51.08	56.75^{c}	83.23^{b}	104.55^{b,d}	123.5^{c}	100.37^{a,c}	66.77^{d}	32.57^{c,d}	19.47^{a}	16.17^{a}
SWAT	3.72	5.47	29.21^{c}	42.72^{a}	63.08^{a}	97.71^{b,c,d}	88.40^{e}	69.87	55.80^{a,b,c}	32.86^{c}	19.47^{a}	7.60^{d,e}
Average ensemble	11.83^{c}	16.17	26.39^{d}	43.90^{a}	69.99	97.94^{d}	115.59^{d}	99.25^{c}	61.03^{b}	30.02^{d}	15.44	9.41^{e}

Note: SWAT, Soil and Water Assessment Tool.
[1]Note that no ETa values were provided for TerraClimate for the months of January and February.

agricultural, forest, urban, and wetland regions, respectively, with clusters identified with superscript letters. Meanwhile, Tables S5–S17 show the same analysis for each individual landuse.

Overall, similar results to the overall analyses were seen, with the winter months, such as January and February, having fewer clusters and more unique datasets, whereas the summer months such as June and July, had more clusters and fewer unique datasets. However, analysis of the major landuse classifications showed that the presence of vegetation might result in less ETa dataset viability.

Seasonal Analysis. Watershed Scale Analysis. Temporal cluster analysis was also performed on a seasonal basis to determine if additional consistent similarities and differences among the ETa datasets could be identified. Table 3 presents the mean monthly seasonal values of each ETa dataset for the entire Honeyoey watershed as well as any similarities between datasets with superscripted letters. The results were similar to the monthly analysis, in which the winter and the fall show fewer clusters and more unique datasets, whereas the spring and the summer show more clusters and fewer unique datasets. This is likely due to the challenges such as cloud cover and snow cover that occur during the fall and the winter seasons (Wang et al. 2015). Meanwhile, unlike the monthly analysis, the average ensemble showed similarities for all seasons, though again no noticeable pattern was seen in which datasets were found to be similar.

Landuse Analysis. Similar to the monthly analysis, major landuse and individual landuse seasonal analysis was performed to determine if consistent similarities and differences among the ETa datasets were lost due to aggregation at the watershed scale. Tables S18–S21 show the mean seasonal values of each ETa dataset for agricultural, forest, urban, and wetland regions, respectively, with clusters identified with super script letters. Meanwhile, Tables S22–S34 show the same analysis for each individual landuse.

TABLE 3. Average seasonal ETa values for each dataset for the entire watershed with clusters indicated by superscripts for each column.

	Seasons			
Datasets	Winter	Spring	Summer	Fall
MOD16A2 1 km	17.65	45.02^{a}	89.02^{a}	29.84
MOD16A2 500 m	15.89	53.49	116.88	33.66^{a}
SSEBop	0.25^{a}	29.07^{b}	103.42^{b,c,d}	23.42^{b}
NLDAS-2:Mosaic	11.51^{b}	60.38^{c}	123.28	52.02
NLDAS-2:Noah	10.03	30.44^{b}	92.26^{a}	35.35^{a,c}
NLDAS-2:VIC	8.17	24.56	101.17^{b,c}	24.30^{b}
TerraClimate	1.09^{a}	68.6^{d}	98.55^{a,b,c,d}	45.76
ALEXI	25.48	63.68^{c,d}	109.47	39.60
SWAT	5.6	45.00^{a}	85.33^{a}	36.04^{a,c}
Average ensemble	12.47^{b}	46.76^{a}	104.26^{c,d}	35.50^{c}

Overall, the seasonal analysis showed similar results to the monthly analysis. This confirmed that the presence of vegetation plays a major role in the similarity between ETa datasets. Furthermore, the spring and summer tended to show more similarities among the ETa datasets, whereas the fall and the winter tended to have fewer similarities and more unique ETa datasets. This matches the weather patterns found in the region and confirms that cloud cover and snow played a major role in ETa dataset variance. Meanwhile, there did not seem to be any noticeable patterns among ETa dataset similarities between the months and seasons. This is likely due to the various accuracies and spatial resolutions associated with the individual ETa products. However, across the different landuses, ETa datasets tended to show similar patterns for specific months and seasons. This is likely due to similarities in how the different ETa datasets were calculated, e.g., Noah, Mosaic, SSEBop, and ALEXI all utilize forms of surface energy balances to calculate ETa, whereas MOD16A2 and SWAT utilize Penman–Monteith techniques, and VIC and TerraClimate utilize water balances.

Spatial Statistical Analysis

Spatial statistical analysis was performed on the ETa datasets, average ensemble, and SWAT model outputs to determine how the different datasets performed across different landuses. The first step in this analysis was to determine how the individual datasets performed across the different landuses found in the study area. After this, a comparison among the different datasets was performed to determine if any of the datasets showed similarities across different landuses.

Landuse Distinction within each ETa Dataset. Overview of Landuse Distinction within each ETa Dataset. Table 4 presents the overall datasets averages with respect to the major landuse categories as well as the watershed scale average. Similarities between these regions are indicated with superscript letters. Overall, a number of similarities were identified. However, these could be caused by similarities in the landuses due to changes throughout the year. Therefore, seasonal and monthly analysis was considered to further examine these similarities and determine their cause.

Seasonal Overview of Landuse Distinction within each ETa Dataset. To determine if the similarities noticed in the watershed scale analysis were related to specific times of the year, seasonal analysis was performed for each dataset, with the results presented in Tables S35–S44. This analysis was also performed on a monthly basis for the major landuse classes (Tables S45–S54) and for individual landuses (Tables S55–S64) with the results presented in Supporting Information.

A summary of this analysis for all of the ETa products can be seen in Table 5. Overall, breaking the spatial analysis down to the seasonal scale improved the individual ETa product performances. This also allowed for the identification of the best seasons to use each dataset: MOD16A2 1 km: spring and summer; MOD16A2 500 m: winter, spring, summer, fall; SSEBop: winter; NLDAS-2 Mosaic: spring and summer; NLDAS-2 Noah: winter and summer; NLDAS-2 VIC: spring; TerraClimate: spring; ALEXI: none; SWAT: spring and summer; Average Ensemble: winter, summer, and fall. This shows that the best product for landuse distinction in the Honeyoey watershed was the MOD16A2 500 m, whereas the worst for the region was the ALEXI product. Furthermore, analysis of the average ensemble showed that by averaging the ETa products it was possible to improve the performance and differentiate among the major landuses.

Landuse Similarities between ETa Datasets. Statistical analysis comparing the ETa datasets was also performed regarding the entire watershed and the major landuse categories (Table 6) and for all individual landuses (Table S65). This is similar to the temporal analysis and was performed to determine if any similarities existed between the ETa datasets when considering the spatial distribution of landuses throughout the region. As can be seen in Table 6, a variety of datasets clusters were identified. However, unlike the temporal analysis, several of the clusters spanned multiple landuses. For example, the MOD16A2 1 km, NLDAS-2:Noah, NLDAS-2:VIC, and SWAT model datasets had similar ETa values across agricultural, urban, and wetland regions. This is interesting since, each of these datasets has different accuracies and spatial and temporal resolutions. However, the governing equations for each of these products is based on energy balances, which could explain why they produced similar values across these landuses. Meanwhile, the ALEXI and TerraClimate and the ALEXI and NLDAS-2: Mosaic ETa products were similar for all major

TABLE 4. Overall dataset averages for each major landuse category with clusters indicated by superscripts for each column.

	Dataset									
Landuse	MOD16A2 1 km	MOD16A2 500 m	SSEBop	NLDAS-2: Mosaic	NLDAS-2: Noah	NLDAS-2: VIC	Terra Climate	ALEXI	SWAT	Average ensemble
Agriculture	43.92	52.60	37.89	62.60^{a}	42.68^{a}	40.14^{a}	66.72^{a}	59.66^{a}	$43.50^{a,b}$	49.39
Forest	47.81^{a}	59.30	40.89^{a}	$61.09^{a,b}$	$40.71^{a,b,c}$	37.24	$66.61^{a,b}$	59.77^{a}	41.45^{a}	50.30^{a}
Urban	41.68	49.48	35.40	$63.28^{a,b}$	43.89^{b}	41.30^{b}	67.01^{b}	56.83	34.25	48.46
Wetland	47.58^{a}	58.00	40.97^{a}	59.85	41.29^{c}	$40.37^{a,b}$	66.70^{a}	59.39^{a}	44.68^{b}	50.39^{a}

TABLE 5. Summary of landuse and season differentiation for all ETa products used in this study, Xs mark conditions that could be differentiated by the product.

Dataset	Spatial scale	Temporal scale				
		Overall	Winter	Spring	Summer	Fall
MOD16A2 1 km	Agriculture	X	X	X	X	X
	Forest			X	X	
	Urban	X		X	X	X
	Wetland		X	X	X	
MOD16A2 500 m	Agriculture	X	X	X	X	X
	Forest	X	X	X	X	X
	Urban	X	X	X	X	X
	Wetland	X	X	X	X	X
SSEBop	Agriculture	X	X		X	X
	Forest		X			
	Urban	X	X		X	X
	Wetland		X			
NLDAS-2: Mosaic	Agriculture			X	X	
	Forest		X	X	X	X
	Urban		X	X	X	X
	Wetland	X		X	X	
NLDAS-2: Noah	Agriculture		X		X	X
	Forest		X		X	X
	Urban		X	X	X	
	Wetland		X		X	
NLDAS-2: VIC	Agriculture		X	X		
	Forest	X		X	X	
	Urban		X	X	X	
	Wetland			X		X
TerraClimate	Agriculture			X	X	
	Forest			X	X	X
	Urban			X		X
	Wetland			X		
ALEXI	Agriculture		X		X	
	Forest					
	Urban	X		X	X	X
	Wetland					X
SWAT	Agriculture			X	X	
	Forest		X	X	X	X
	Urban	X		X	X	
	Wetland		X	X	X	X
Average ensemble	Agriculture	X	X		X	X
	Forest		X		X	X
	Urban	X	X		X	X
	Wetland		X		X	X

landuses. The similarity between the ALEXI and Mosaic products can also be explained by the fact that both use energy balances to calculate ETa, which is similar the first cluster of ETa products. However, considering the TerraClimate/ALEXI similarity, these products utilize different governing equations (water balance vs. energy balance) and have very different accuracies. However, they share the same spatial resolution. This in combination with the other clusters indicates that ETa product similarity can be obtained for datasets that utilize similar approaches; however, it is also possible to achieve this same result if datasets share a spatial resolution. However, given the spatial nature of ETa this makes sense.

On the other hand, at the watershed scale, there was a reduction in the number of datasets sets found to be similar, e.g., at the watershed scale the MOD16A2 1 km dataset was not considered similar to the NLDAS-2: Noah, NLDAS-2: VIC, and SWAT datasets, which was seen when considering specific landuses. This shows that aggregation to the watershed level can result in the loss of similarities among ETa products, which indicates that performing analysis for specific landuses improves overall product agreement.

TABLE 6. Overall summary of average ETa values for each dataset for the entire watershed and each major landuse category with clusters indicated by superscripts for each column.

Dataset	Region: Watershed	Agricultural	Forest	Urban	Wetlands
MOD16A2 1 km	45.39^{a}	43.92^{a}	47.82^{a}	41.68^{a}	$47.59^{a,b}$
MOD16A2 500 m	54.98^{b}	52.61^{b}	$59.30^{b,c}$	49.49^{b}	$58.00^{c,d}$
SSEBop	39.05^{c}	37.89	40.90^{d}	35.40^{c}	40.98^{a}
NLDAS-2:Mosaic	61.80^{d}	62.61^{c}	61.09^{b}	63.28^{d}	59.86^{c}
NLDAS-2:Noah	$42.03^{a,c}$	42.68^{a}	40.72^{d}	43.89^{a}	41.30^{a}
NLDAS-2:VIC	39.56^{c}	40.14^{a}	37.24^{e}	$41.30^{a,c}$	40.38^{a}
TerraClimate	$66.71^{b,d}$	$66.73^{b,c}$	66.62^{c}	67.01^{e}	66.70^{d}
ALEXI	$59.57^{b,d}$	$59.67^{b,c}$	$59.77^{b,c}$	$56.84^{d,e}$	$59.39^{c,d}$
SWAT	43.00^{c}	43.50^{a}	$41.45^{d,e}$	$34.25^{a,c}$	$44.68^{a,b}$
Average ensemble	49.75	49.39	50.31^{a}	48.46^{b}	50.39^{b}

Overall, the presence of recurring similarities and differences among the ETa products when considering landuses shows that agreement among the ETa products is possible and is influenced by spatially dependent variables such as landuse and governing equations. This makes sense given the spatial variability associated with ETa; however, when combined with the lack of consistent similarities and differences in the temporal analysis indicates that landuse plays a more important role for ETa than seasonal variations, at least with respect to product agreement.

Subbasin-Level Statistical Analysis

SWAT Model Output. Similar to the landuse analysis performed between all of the datasets, spatial mean difference was also calculated between the eight remotely sensed ETa products and the SWAT model output at the subbasin level. This analysis, presented graphically in Figure 3, provides a spatial overview regarding how well the SWAT model was able to replicate the remotely sensed ETa datasets. As can be seen in Figure 3, most of the subbasins in maps a), c), e), and f), MOD16A2 1 km, SSEBop, NLDAS-2: Noah and NLDAS-2: VIC, respectively, show no difference in their ETa values with the SWAT model output. This matches the results seen in Landuse Similarities between ETa Datasets section, which also showed that these datasets were more closely aligned with the SWAT model regarding individual landuse ETa values. However, while the previous analysis indicated that these datasets were similar in nature, the spatial subbasin analysis shows that the SWAT model output is over or underestimating different regions within the Honeyoey watershed. This shows that while the watershed scale analysis showed agreement, it is important to take into account spatial variation within the landscape. On the other hand, when considering the other remotely sensed ETa datasets, most of the region shows that the SWAT model is underestimating the values reported by the MOD16A2 500 m, NLDAS-2: Mosaic, TerraClimate, and ALEXI products. This also matches the earlier results of this study as well as the results of previous studies that showed that the SWAT model had a better fit with the SSEBop dataset compared to the ALEXI dataset (Herman et al. 2018). These results are further supported by subbasin level statistical difference/no difference presented in Figure S1.

Average Ensemble. Subbasin level analysis was also performed comparing the average ensemble's ETa values to the eight remotely sensed ETa products and the SWAT model output (Figure 4). As can be seen in Figure 4, the average ensemble was either under- or overestimating ETa values for all of the datasets. Interestingly the split as to which datasets was under- or overestimated matched the split of those datasets that were either similar of different from the SWAT model output. With MOD16A2 1 km, SSEBop, NLDAS-2: Noah and NLDAS-2: VIC showing that the average ensemble overestimated Eta values, whereas comparisons to MOD16A2 500 m, NLDAS-2: Mosaic, TerraClimate, and ALEXI products showed underestimation. However, this makes sense since the average ensemble was created by averaging all datasets used in this study. This would result in a dataset that fits the middle ground between all datasets, which is clearly the case here. This also explains why the average ensemble was found to be statistically different for the majority of subbasins for all datasets (Figure S2).

Time Series Analysis

This section focuses mainly on the temporal variabilities of the remotely sensed and SWAT ETa

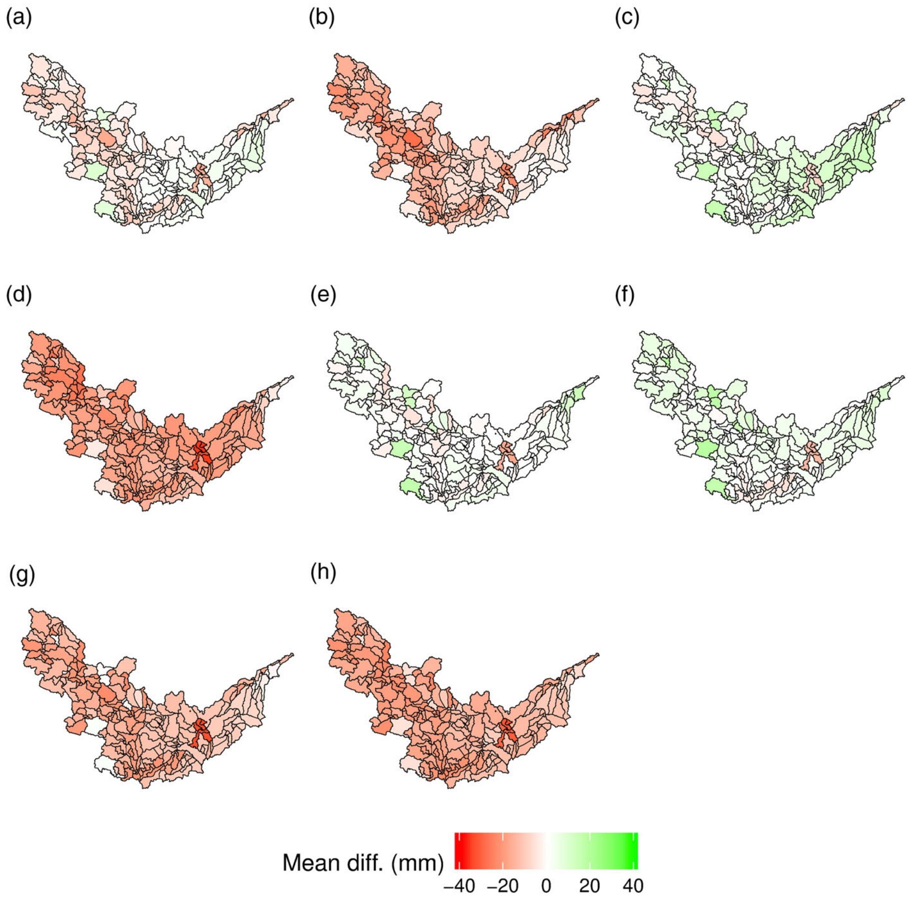

FIGURE 3. Maps showing the mean difference between each ETa dataset and the SWAT model output. Maps correspond to (a) MOD16A2 1 km, (b) MOD16A2 500 m, (c) SSEBop, (d) NLDAS-2: Mosaic, (e) NLDAS-2: Noah, (f) NLDAS-2: VIC, (g) TerraClimate, and (h) ALEXI.

products against the average ensemble ETa. Figure 5 presents the comparison between the RS ETa monthly time series at the watershed level against the Ensemble, from 2003 to 2014. In general, SSEBop, TerraClimate and NLDAS2 (i.e., Mosaic, Noah, and VIC) products provide minimum ETa values below the minimum ensemble values. Meanwhile, MOD16A2 500 m, Mosaic, and ALEXI provide maximum values above the Ensemble maximum values. However, in order to better capture these variabilities on the on a monthly basis, Figure S3 was created.

Figure S3 shows the monthly distribution of the difference between RS ETa datasets and the Ensemble ($ETa_{dataset}$–$Ta_{Ensemble}$) at the watershed level. Overall, SSEBop, MOD16A2 500 m, NLDAS2 Mosaic, and NLDAS2 VIP are overpredicting ETa values for the months of June, July, and August. This makes sense since these months are the hottest months of the year. However, higher ETa rates were observed for the coldest months of the year (January and February) by MOD16A2 1 km and ALEXI, which can indicate the lower reliability of these products in colder times of the year. It is also worth noting that the TerraClimate dataset results in the highest magnitude and variability of these differences. This can represent that the ETa product of TerraClimate is fundamentally different from the majority of ETa RS products used in this study since it uses a water balance approach to calculating ETa. Figure S4 presents the comparison between the average ensemble and the SWAT simulation for both the watershed level ETa monthly time series and the monthly distribution of ETa differences. This figure indicates that SWAT generally provides lower ETa values than the Ensemble. However, the highest divisions were observed in July and August that can be significant for underestimating crop water requirements. Seasonally, the highest differences between different ETa products and the average ensemble are provided by ALEXI for the winter, TerraClimate for the spring, and Mosaic for the summer and the fall (see Figure S5). Likewise, the lowest differences are given by SSEBop for the winter and fall, Noah for the spring, and TerraClimate and SWAT for the summer. In general, the lowest variabilities can be a good indicator of the reliability of the dataset when no

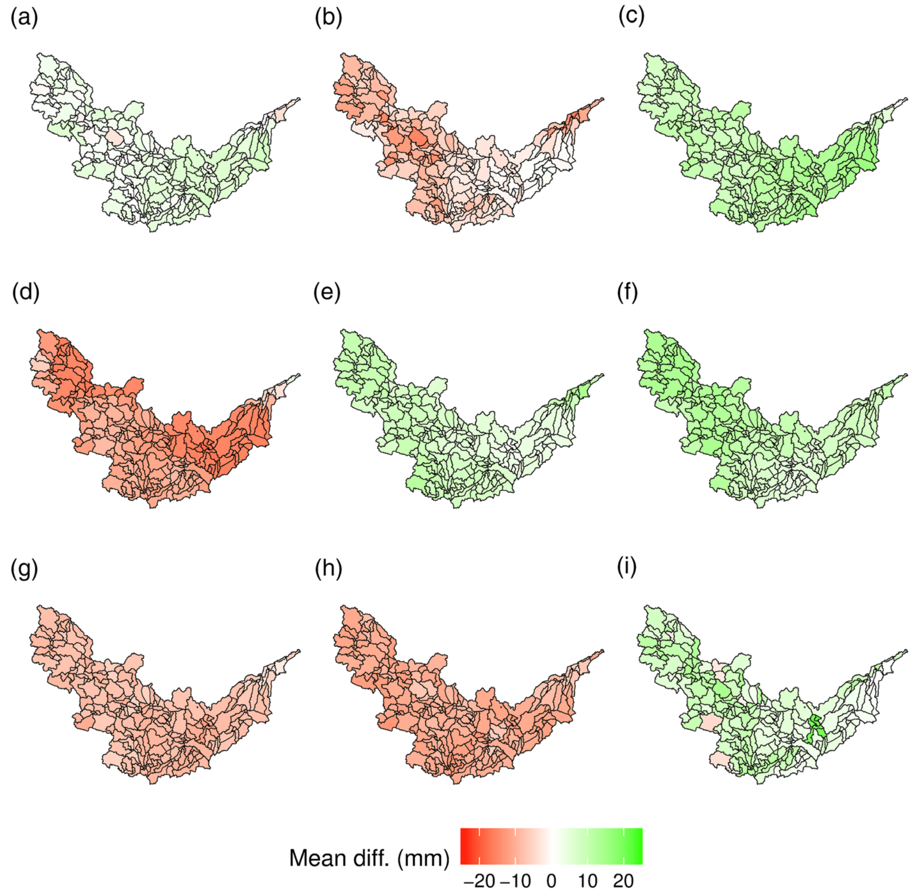

FIGURE 4. Maps showing the mean difference between each ETa dataset and the average ensemble. Maps correspond to (a) MOD16A2 1 km, (b) MOD16A2 500 m, (c) SSEBop, (d) NLDAS-2: Mosaic, (e) NLDAS-2: Noah, (f) NLDAS-2: VIC, (g) TerraClimate, (h) ALEXI, and (i) SWAT model.

additional information exists regarding validating the accuracy of the ETa product for the region of study.

CONCLUSIONS

Throughout the course of this study statistical analysis was used to compare the performance of published remotely sensed ETa products in a region with no observed ETa data. Overall, temporal analysis of the datasets showed that there was no noticeable trend in similarities between specific datasets at both monthly and seasonal scales. However, a general pattern was seen with the summer and spring seasons showing more clusters among the datasets and fewer unique datasets. Meanwhile, the fall and the winter seasons showed fewer clusters and more unique ETa datasets. This is reflective of weather and vegetation cover trends within the region. Nevertheless, the lack of consistent similarities and differences among the datasets shows that temporal variation is less influential when compared to spatial variation. This likely due to several factors such as the spatial resolutions. This was most clearly identified in the comparison of the two MOD16A2 products (1 km and 500 m). Despite both products utilizing the same approach and temporal resolution, the lack of similarity throughout the year can be attributed to the impact of different spatial resolutions.

Meanwhile, spatial analysis at both the watershed, landuse, and subbasin levels led to the identification of two major clusters within the ETa datasets. These clusters were consistent across different landuses. This highlights two major points, first, there is lots of variance among the different RS ETa products, which is driven by the use of *different* governing equations, spatial and temporal resolutions, and accuracies. However, the second point is that it is possible to find similar ETa time series across different RS ETa products, this is driven by the use of *similar* governing equations and spatial resolutions.

In addition, it is important to note which datasets should be used when. The ETa product that was able to differentiate among all of the major landuses for

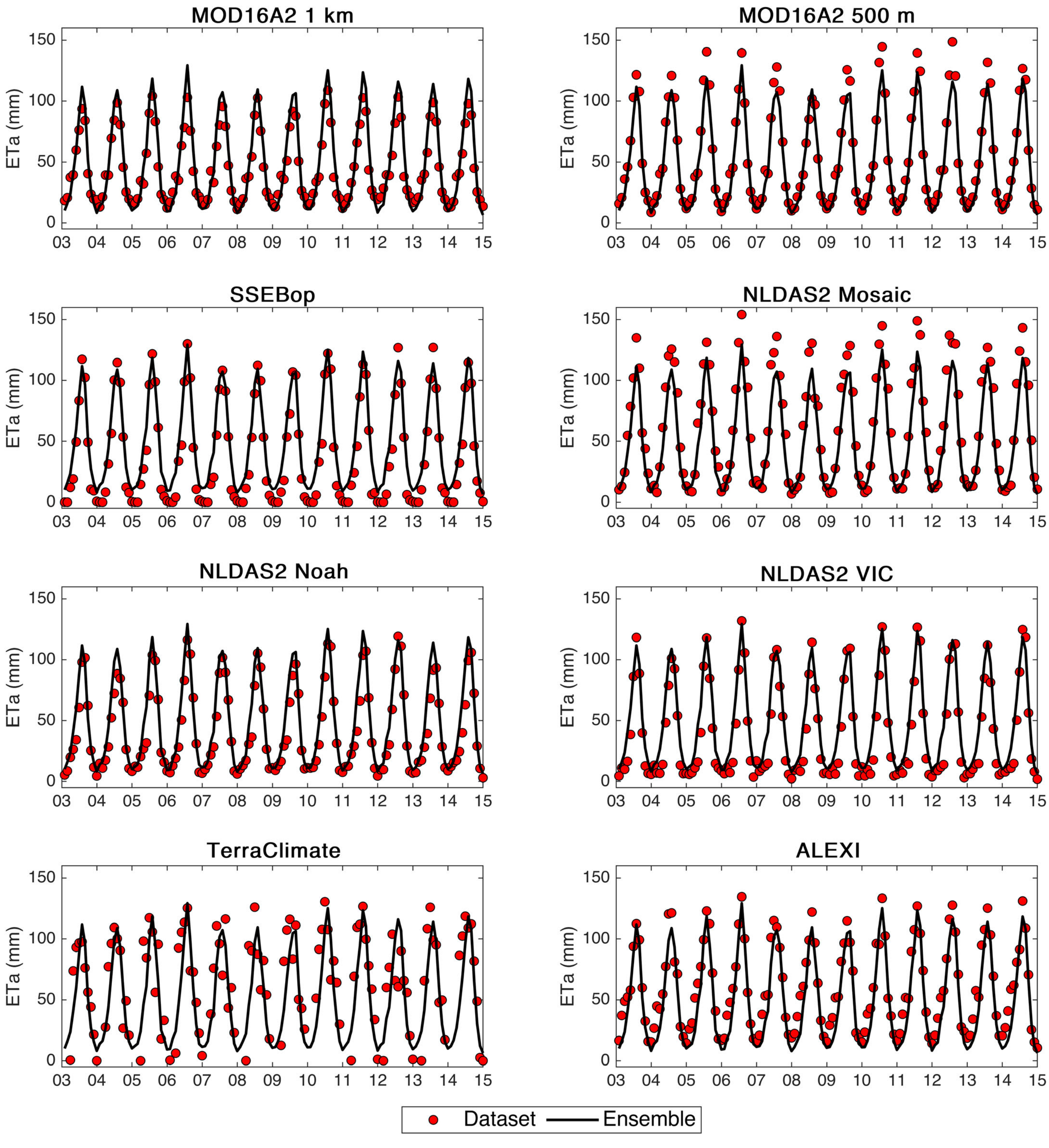

FIGURE 5. Monthly ETa time series aggregated at the watershed level for different remote sensing-based products compared with the Ensemble product from 2003 to 2015.

all seasons was the MOD16A2 500 m dataset. However, all of the other datasets, except for ALEXI, were able to differentiate between landuses for at least one season.

In summary, this study showed that remotely sensed ETa products do not follow the same trend throughout the year, this is more pronounced during the period of low vegetation cover. However, we identified the appropriate application period for different ETa products. Finally, spatial variabilities of the ETa products are more in tune with landuse and climate characteristics and can be explained based on the governing equations, spatial and temporal resolutions, and accuracies of the products. This can also help stakeholders, policy makers, and researchers select the best ETa dataset for different tasks such as the monitoring of agricultural lands or tracking deforestation.

However, this study was performed for only one watershed, future studies should be performed to expand this analysis to different climatological zones. This would help improve our understanding of how each ETa product performs across the global landscape and which ones should or should not be recommended for different seasons or landuses to help ensure that the correct ETa dataset is selected. Furthermore, other ensembling techniques should be performed to identify the best product for different regions.

SUPPORTING INFORMATION

Additional supporting information may be found online under the Supporting Information tab for this article: Supporting Information includes additional tables and figures used in temporal and spatial analysis among the ETa datasets performed in this paper.

ACKNOWLEDGMENTS

The authors thank Dr. Martha C. Anderson from USDA-ARS Hydrology and Remote Sensing Laboratory at Beltsville, Maryland and Dr. Christopher R. Hain from NASA Marshall Space Flight Center at Huntsville, AL for his help in providing ALEXI data. This work is supported by the USDA National Institute of Food and Agriculture, Hatch project MICL02653.

LITERATURE CITED

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." *Scientific Data* 5: 170191.

AmeriFlux. 2018. "AmeriFlux." http://ameriflux.lbl.gov/.

Anderson, M.C., R.G. Allen, A. Morse, and W.P. Kustas. 2012. "Use of Landsat Thermal Imagery in Monitoring Evapotranspiration and Managing Water Resources." *Remote Sensing of Environment* 122: 50–65.

Anderson, M.C., J.M. Norman, G.R. Diak, W.P. Kustas, and J.R. Mecikalski. 1997. "A Two-Source Time-Integrated Model for Estimating Surface Fluxes Using Thermal Infrared Remote Sensing." *Remote Sensing of Environment* 60 (2): 195–216.

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." *Journal of Geophysical Research: Atmospheres* 112 (D10). https://doi.org/10.1029/2006JD007506

Bhattarai, N., S.B. Shaw, L.J. Quackenbush, J. Im, and R. Niraula. 2016. "Evaluating Five Remote Sensing Based Single-Source Surface Energy Balance Models for Estimating Daily Evapotranspiration in a Humid Subtropical Climate." *International Journal of Applied Earth Observation and Geoinformation* 49: 75–86.

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." *Agricultural and Forest Meteorology* 186: 1–11.

Christensen, N., and D.P. Lettenmaier. 2006. "A Multimodel Ensemble Approach to Assessment of Climate Change Impacts on the Hydrology and Water Resources of the Colorado River Basin." *Hydrology and Earth System Sciences Discussions* 3 (6): 3727–70.

Clark, M.P., Y. Fan, D.M. Lawrence, J.C. Adam, D. Bolster, D.J. Gochis, R.P. Hooper et al. 2015. "Improving the Representation of Hydrologic Processes in Earth System Models." *Water Resources Research* 51 (8): 5929–56.

Dietterich, T.G.2000. "Ensemble Methods in Machine Learning." In International Workshop on Multiple Classifier Systems (1–15), Springer, Berlin, Heidelberg.

Duan, Q., N.K. Ajami, X. Gao, and S. Sorooshian. 2007. "Multi-Model Ensemble Hydrologic Prediction Using Bayesian Model Averaging." *Advances in Water Resources* 30 (5): 1371–86.

Enviroweather. (2018). "Enviroweather: Weather-Based Pest, Natural Resources, and Production Management Tools." https://www.enviroweather.msu.edu/.

EPA. 2017. "Saginaw River and Bay Area of Concern." https://19january2017snapshot.epa.gov/saginaw-river-bay-aoc_.html.

Ershadi, A., M.F. McCabe, J.P. Evans, N.W. Chaney, and E.F. Wood. 2014. "Multi-Site Evaluation of Terrestrial Evaporation Models Using FLUXNET Data." *Agricultural and Forest Meteorology* 187: 46–61.

Ershadi, A., M.F. McCabe, J.P. Evans, and E.F. Wood. 2015. "Impact of Model Structure and Parameterization on Penman-Monteith Type Evaporation Models." *Journal of Hydrology* 525: 521–35.

FAA. 2018. "Surface Weather Observation Stations: ASOS/AWOS." https://www.faa.gov/air_traffic/weather/asos/?state=MI.

Fowler, H.J., and M. Ekström. 2009. "Multi-Model Ensemble Estimates of Climate Change Impacts on UK Seasonal Precipitation Extremes." *International Journal of Climatology* 29 (3): 385–416.

Fox, J., and G. Monette. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks, CA: Sage.

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." *Transactions of the American Society of Agricultural and Biological Engineers* 50 (4): 1211–50.

Giorgi, F., and L.O. Mearns. 2003. "Probability of Regional Climate Change Based on the Reliability Ensemble Averaging (REA) Method." *Geophysical Research Letters* 30 (12). https://doi.org/10.1029/2003GL017130.

Giri, S., A.P. Nejadhashemi, and S.A. Woznicki. 2012. "Evaluation of Targeting Methods for Implementation of Best Management

Practices in the Saginaw River Watershed." *Journal of Environmental Management* 103: 24–40.
Google. 2018. "EEFlux." https://eeflux-level1.appspot.com/.
Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment* 202: 18–27.
Graham, S. 1999. *Remote Sensing*. NASA Earth Observatory. http://earthobservatory.nasa.gov/Features/RemoteSensing/printall.php.
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." *Journal of Hydrometeorology* 16 (3): 1070–86.
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." *Journal of Hydrology* 556: 39–49.
Herman, M.R., A.P. Nejadhashemi, F. Daneshvar, D.M. Ross, S.A. Woznicki, Z. Zhang, and A.H. Esfahanian. 2015. "Optimization of Conservation Practice Implementation Strategies in the Context of Stream Health." *Ecological Engineering* 84: 1–12.
Kim, J., B.P. Mohanty, and Y. Shin. 2015. "Effective Soil Moisture Estimate and Its Uncertainty Using Multimodel Simulation Based on Bayesian Model Averaging." *Journal of Geophysical Research: Atmospheres* 120 (16): 8023–42.
Kite, G.W., and P. Droogers. 2000. "Comparing Evapotranspiration Estimates from Satellites, Hydrological Models and Field Data." *Journal of Hydrology* 229 (1–2): 3–18.
Krishnamurti, T.N., C.M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran. 2000. "Multimodel Ensemble Forecasts for Weather and Seasonal Climate." *Journal of Climate* 13 (23): 4196–216.
Kustas, W.P., and J.M. Norman. 2000. "A Two-Source Energy Balance Approach Using Directional Radiometric Temperature Observations for Sparse Canopy Covered Surfaces." *Agronomy Journal* 92 (5): 847–54.
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 AGU Fall Meeting Abstracts.
Long, D., L. Longuevergne, and B.R. Scanlon. 2014. "Uncertainty in Evapotranspiration from Land Surface Modeling, Remote Sensing, and GRACE Satellites." *Water Resources Research* 50 (2): 1131–51.
Love, B., and A.P. Nejadhashemi. 2011. "Environmental Impact Analysis of Biofuel Crops Expansion in the Saginaw River Watershed." *Journal of Biobased Materials and Bioenergy* 5 (1): 30–54.
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." *Journal of Geophysical Research: Atmospheres* 123 (2): 814–34.
Martinez-Martinez, E., A.P. Nejadhashemi, S.A. Woznicki, and B.J. Love. 2014. "Modeling the Hydrological Significance of Wetland Restoration Scenarios." *Journal of Environmental Management* 133: 121–34.
McVicar, T.R., M.L. Roderick, R.J. Donohue, L.T. Li, T.G. Van Niel, A. Thomas, J. Grieser et al. 2012. "Global Review and Synthesis of Trends In Observed Terrestrial Near-Surface Wind Speeds: Implications for Evaporation." *Journal of Hydrology* 416: 182–205.
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." *Transactions of the American Society of Agricultural and Biological Engineers* 50 (3): 885–900.
Mu, Q., F.A. Heinsch, M. Zhao, and S.W. Running. 2007. "Development of a Global Evapotranspiration Algorithm Based on MODIS and Global Meteorology Data." *Remote sensing of Environment* 111 (4): 519–36.
Mu, Q., M. Zhao, and S.W. Running. 2011. "Improvements to a MODIS Global Terrestrial Evapotranspiration Algorithm." *Remote Sensing of Environment* 115 (8): 1781–800.
Mueller, B., S.I. Seneviratne, C. Jimenez, T. Corti, M. Hirschi, G. Balsamo, P. Ciais et al. 2011. "Evaluation of Global Observations-Based Evapotranspiration Datasets and IPCC AR4 Simulations." *Geophysical Research Letters* 38 (6). https://doi.org/10.1029/2010GL046230.
NASA. 2014. "MOD16A2: MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500 m SIN Grid V006." https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod16a2_v006.
NASA. 2018a. "MOD 16 — Evapotranspiration." https://modis.gsfc.nasa.gov/data/dataprod/dataproducts.php?MOD_NUMBER=16.
NASA. 2018b. "MOD16A2 — MODIS/Terra Evapotranspiration 8-Day Level-4 Global 500m SIN Grid." https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/evapotranspiration/MOD16A2/.
NASA. 2018c. "NLDAS Concept/Goals." https://ldas.gsfc.nasa.gov/nldas/NLDASgoals.php.
NASA/GSFG. 2018. "Goddard Earth Science Data and Information Services Center (GES DISC)." https://disc.gsfc.nasa.gov/.
NASS. 2018. "CropScape — Cropland Data Layer." https://nassgeodata.gmu.edu/CropScape/.
NCDC. 2018. "Data Tools: Local Climatological Data (LCD)." https://www.ncdc.noaa.gov/cdo-web/datatools/lcd
Neitsch, S.L., J.G. Arnold, J.R. Kiniry, and J.R. Williams. 2011. *Soil and Water Assessment Tool Theoretical Documentation Version 2009*. Texas Water Resources Institute. http://hdl.handle.net/1969.1/128050.
Nejadhashemi, A.P., B.J. Wardynski, and J.D. Munoz. 2012. "Large-Scale Hydrologic Modeling of the Michigan and Wisconsin Agricultural Regions to Study Impacts of Land Use Changes." *Transactions of the American Society of Agricultural and Biological Engineers* 55 (3): 821–38.
NOAA. 2017. "Potential Evapotranspiration." https://www.ncdc.noaa.gov/monitoring-references/dyk/potential-evapotranspiration.
NOAA-NCEP. 2013. "Climate Prediction Center — wgrib Home Page." http://www.cpc.ncep.noaa.gov/products/wesley/wgrib.html.
NRCS. 2018. "Web Soil Survey." https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx.
NTSG. 2018. "MODIS Global Evapotranspiration Project (MOD16)." 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.
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." *Journal of the American Water Resources Association* 49 (3): 577–91.
Singh, R., and G. Senay. 2016. "Comparison of Four Different Energy Balance Models for Estimating Evapotranspiration in the Midwestern United States." *Water* 8 (1): 9.
Srinivasan, V., M. Konar, and M. Sivapalan. 2017. "A Dynamic Framework for Water Security." *Water Security* 1: 12–20.
Tebaldi, C., and R. Knutti. 2007. "The Use of the Multi-Model Ensemble in Probabilistic Climate Projections." *Philosophical*

Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 365 (1857): 2053–75.

Texas A&M University. 2018. "SWAT: Soil and Water Assessment Tool." 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 AGU Fall Meeting Abstracts.

van der Tol, C., and G.N. Parodi. 2012. "Guidelines for Remote Sensing of Evapotranspiration." In *Evapotranspiration-Remote Sensing and Modeling*, edited by A. Irmak, 227–50. Rijeka: InTech.

U.S. Climate Data. 2018. "Climate Mount Pleasant — Michigan." U.S. Climate Data. https://www.usclimatedata.com/climate/mount-pleasant/michigan/united-states/usmi0577/2018/1.

USGS. 2016a. "Evapotranspiration — The Water Cycle." USGS Water Science School. https://water.usgs.gov/edu/watercycleevapotranspiration.html.

USGS. 2016b. "The Water Cycle: Evaporation." USGS Water Science School. https://water.usgs.gov/edu/watercycleevaporation.html.

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 Sensing of Environment* 139: 35–49.

Vinukollu, R.K., E.F. Wood, C.R. Ferguson, and J.B. Fisher. 2011. "Global Estimates of Evapotranspiration for Climate Studies Using Multi-Sensor Remote Sensing Data: Evaluation of Three Process-Based Approaches." *Remote Sensing of Environment* 115 (3): 801–23.

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 Resources Research* 50 (8): 6874–91.

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 Resources Research* 54 (3): 2129–51.

Wang, S., M. Pan, Q. Mu, X. Shi, J. Mao, C. Brümmer, R.S. Jassal, P. Krishnan, J. Li, and T.A. Black. 2015. "Comparing Evapotranspiration from Eddy Covariance Measurements, Water Budgets, Remote Sensing, and Land Surface Models over Canada." *Journal of Hydrometeorology* 16 (4): 1540–60.

Wang, Z. 2013. "cts: An R Package for Continuous Time Autoregressive Models via Kalman Filter." *Journal of Statistical Software* 53 (5): 1–19.

Wu, B.F., J. Xiong, N.N. Yan, L.D. Yang, and X. Du. 2008. "ETWatch for Monitoring Regional Evapotranspiration with Remote Sensing." *Advances in Water Science* 19 (5): 671–78.

Xia, Y., M.T. Hobbins, Q. Mu, and M.B. Ek. 2015. "Evaluation of NLDAS-2 Evapotranspiration Against Tower Flux Site Observations." *Hydrological processes* 29 (7): 1757–71.

Yao, Y., S. Liang, X. Li, J. Chen, S. Liu, K. Jia, X. Zhang et al. 2017. "Improving Global Terrestrial Evapotranspiration Estimation Using Support Vector Machine by Integrating Three Process-Based Algorithms." *Agricultural and Forest Meteorology* 242: 55–74.

Zhang, K., J.S. Kimball, and S.W. Running. 2016. "A Review of Remote Sensing Based Actual Evapotranspiration Estimation." *Wiley Interdisciplinary Reviews: Water* 3 (6): 834–53.