Published in final edited form as:
J Hydrometeorol. 2018 May ; 19(5): 891–905. doi:10.1175/jhm-d-17-0186.1.

A water balance based, spatiotemporal evaluation of terrestrial evapotranspiration products across the contiguous United States

Elizabeth Carter[1,*] **[PhD Candidate]**, **Christopher Hain**[2] **[Research Scientist]**, **Martha Anderson**[3] **[Research Physical Scientist]**, **Scott Steinschneider**[4] **[Assistant Professor]**

[1]Department of Biological and Environmental Engineering, Cornell University, 111 Wing Drive, R, Riley-Robb Hall, Ithaca, NY, 14853

[2]NASA Short-term Prediction Research and Transition Center, 320 Sparkman Drive, Huntsville, AL 35805

[3]USDA-ARS Hydrology and Remote Sensing Laboratory, 104 Building 007, BARC-West, Beltsville, MD 20705

[4]Department of Biological and Environmental Engineering, Cornell University, 111 Wing Drive, Riley-Robb Hall, Ithaca, NY, 14853

Abstract

Accurate gridded estimates of evapotranspiration (ET) are essential to the analysis of terrestrial water budgets. In this study, ET estimates from three gridded energy-balance based products (ET_{EB}) with independent model formations and data forcings are evaluated for their ability to capture long term climatology and inter-annual variability in ET derived from a terrestrial water budget (ET_{WB}) for 671 gaged basins across the CONUS. All three ET_{EB} products have low spatial bias and accurately capture inter-annual variability of ET_{WB} in the central US, where ET_{EB} and ancillary estimates of change in total surface water storage (ΔTWS) from the GRACE satellite project appear to close terrestrial water budgets. In humid regions, ET_{EB} products exhibit higher long-term bias, and the covariability of ET_{EB} and ET_{WB} decreases significantly. Several factors related to either failure of ET_{WB}, such as errors in ΔTWS and precipitation, or failure of ET_{EB}, such as treatment of snowfall and horizontal heat advection, explain some of these discrepancies. These results mirror and build on conclusions from other studies: on inter-annual timescales, ΔTWS and error in precipitation estimates are non-negligible uncertainties in ET estimates based on a terrestrial water budget, and this confounds their comparison to energy balance ET models. However, there is also evidence that in at least some regions, climate and landscape features may also influence the accuracy and long-term bias of ET estimates from energy balance models, and these potential errors should be considered when using these gridded products in hydrologic applications.

[*]*Corresponding author* ekc76@cornell.edu.

1. INTRODUCTION

After precipitation, evapotranspiration (ET) is the largest term in the terrestrial surface hydrologic budget (Haddeland et al. 2011). Virtually all water taken up by vegetation is evapotranspired, and as the largest anthropogenic use of surface and groundwater is the irrigation of cropland (Shiklomanov 2000, Döll et al., 2012), characterization of ET has important implications for managing human water consumption. The importance of ET for surface hydrologic budgets and agricultural water use motivates the need for continuous observation-based data sources that can characterize the spatial and temporal variability of ET at a level of detail that can help resolve key aspects of the terrestrial hydrologic cycle.

Surface energy budget based models of ET have substantially enhanced the availability of spatially and temporally continuous ET estimates. Remotely sensed data have also been integrated into surface energy models (Kalma et al. 2008, Li et al. 2009, Glenn et al. 2011), including land surface models (LSMs, Lo et al. 2015), to further increase the accuracy of energy budget based ET estimates. Several methods for integrating remotely sensed observations into energy budget based ET estimates are available, ranging from using remotely sensed data to increase resolution of vegetation parameters in surface energy models (Mu et al., 2011) to utilization of direct measurements of surface energy fluxes (Kustas et al., 2011).

Regardless of whether remotely sensed data are used, most gridded ET products derived from surface energy budget models (hereafter referred to as ET_{EB} products) still require ancillary data on surface meteorological conditions, including but not limited to incident and reflected short wave radiation, near-surface air temperature, vapor pressure, and wind speed (Kalma et al. 2008, Glenn et al. 2011). Because of differences in modeling approaches with regards to remotely sensed and meteorological data inputs, different ET_{EB} products show unique patterns of space-time bias which are related to propagation of errors in forcing data (Ferguson et al. 2010, Vinukollu et al. 2011, Cai et al. 2014), conditions under which models are not well calibrated (Hain et al. 2015, Ferguson et al. 2010, French et al. 2005, Jiménez et al. 2011, Velpuri et al. 2013, Vinukollu et al. 2011), and to climatic conditions where model assumptions fail (Anderson et al. 2012, Choi et al. 2009, Kustas et al. 2012).

Because of a lack of spatially continuous ground-based measurements of ET, ET_{EB} products are difficult to validate at large spatial scales. One of the most common approaches to validate ET_{EB} estimates is to evaluate their consistency with empirical estimates of ET based on a terrestrial water budget (ET_{WB}), calculated for individual basins as:

$$ET = P + G - Q - \Delta TWS \quad \text{Eq. 1.1}$$

where ET (mm) is the mean basin ET, P (mm) is mean basin precipitation, G (mm) accounts for net inter-basin groundwater flux, Q (mm) is basin runoff (normalized by drainage area), and ΔTWS (mm) is the change in basin total water storage between the current and previous time step (where total water storage is the total water stored in soil, surface water, groundwater, and vegetation). Inter-basin groundwater flux is often assumed to be zero, especially in basins with limited anthropogenic withdrawals. In gaged basins, Q is known

with high accuracy. If estimates of basin P are also available, then a water-budget based estimate of ET (ET_{WB}) can be developed by assuming ΔTWS is zero:

$$ET_{WB} = P - Q \quad \text{Eq. 1.2}$$

The approximation in Eq. 1.2 has conventionally been accepted if all terms are evaluated over a sufficiently long time period (annual or longer) to support the assumption that ΔTWS=0 (Twine et al. 2004).

There has been consensus that ET_{EB} products fail to close long-term (Sheffield et al. 2009, Gao et al. 2009) and inter-annual (Zhang et al. 2012, Zeng et al. 2014, Velpurri et al. 2013, Han et al. 2015, Liu et al. 2016) terrestrial water budgets (i.e., $ET_{EB} \neq ET_{WB}$ in many regions in the contiguous U.S. (CONUS). Long-term non-closure errors ($\overline{\varepsilon_{NC}} = \overline{ET_{EB}} - \overline{ET_{WB}}$, with over-bars indicating multi-year averages) follow distinct spatial and climatic gradients, with arid regions exhibiting substantially lower long-term bias than humid regions (Sheffield et al. 2009, Gao et al. 2009). Similarly, most ET_{EB} estimates appear better suited to estimate both multi-year trends and inter-annual variability in arid regions than in humid regions (Zhang et al. 2012, Zeng et al. 2014, Velpurri et al. 2013). As such, interannual non-closure errors ($\varepsilon_{NC} = ET_{EB} - ET_{WB}$) exhibit similar regional patterns as long-term bias ($\overline{\varepsilon_{NC}}$).

The spatial characterization of long term and inter-annual non-closure errors provides insight into the accuracy of different ET_{EB} products for different regions, providing valuable guidance for how such products can be used in different hydrologic applications. Unfortunately, these error estimates are confounded by potential error embedded in ET_{WB}. For instance, basin-average P can exhibit long-term mean bias if estimated from a sparse gaging network, particularly in basins with stark topographic gradients (Prat et al. 2015). The inter-annual variance of basin-average P can also be biased if the variance of precipitation at the gages used to develop the basin-averaged estimate is a poor representation of the true precipitation variability across the basin. Most likely, the variance of basin-average P will be too high because variances of true areal averages should be lower than individual point estimates, but there are limited numbers of gages available to estimate the areal averages. In addition to precipitation biases, recent studies that have incorporated changes to ΔTWS derived from the Gravity Recovery and Climate Experiment (GRACE) satellites indicate that inter-annual variability in ΔTWS may be an important component of the annual terrestrial water budget in many regions (Rodell and Famiglietti 2001, Rodell et al. 2004, Sahoo et al., 2011, Zeng et al. 2012, Han et al. 2015). Therefore, non-closure errors between ET_{WB} and ET_{EB} are difficult to disentangle from uncertainty in basin P and ΔTWS. There have been attempts to disentangle these terms using sensitivity analyses to elucidate regions where ε_{NC} is significantly influenced by precipitation uncertainty (Ferguson et al., 2009; Sheffield et al. 2009; Gao et al., 2009; Hain et al., 2014), or GRACE Tellus based measures of ΔTWS (Sheffield et al. 2009; Gao et al., 2010), with results generally showing that precipitation uncertainty is an important component of ε_{NC}.

A key question underscoring these efforts is whether non-closure errors are caused by bias in the mean or variance of components used to estimate ET_{WB} (e.g., P, ΔTWS), or if the model formulations that underscore different ET_{EB} products fail to capture fundamental processes that drive ET variability. In this study, we seek to advance the evaluation of ET_{EB} across the CONUS, building on recent work exploring the accuracy of ET_{EB} estimates at different spatial and temporal scales (Zhang et al., 2012; Han et al., 2015). Han et al. (2015) explored evidence of hydrologic consistency between ET_{WB} and two very different ET_{EB} products: the NOAH land surface model (Chen et al., 1996) and the ALEXI model (Anderson et al. 1997). In that study, the authors hypothesized that inter-annual non-closure errors in NOAH ET estimates $\left(ET_{EB}^{NOAH}\right)$ were not due to failure of the model as previously argued (Zhang et al., 2012), but rather were linked to errors in estimated values of ET_{WB}. They found a marked improvement in correlation between ET_{WB} and ET_{EB}^{NOAH} over a 10 year period in the Mississippi River Basin, particularly in the Ohio and Upper Mississippi sub-basins, when they included GRACE Tellus ΔTWS in the ET_{WB} approximation in Eq. 1.1 (assuming G to equal zero). Further, they found that the inter-annual correlation between ET_{EB}^{NOAH} and ALEXI ET estimates $\left(ET_{EB}^{ALEXI}\right)$ was high across 15 different basins in the central U.S., although the magnitude of inter-annual variability of ET from both products was low compared to the inter-annual variability of ET_{WB}. They concluded that higher inter-annual variability in ET_{WB} over ET_{EB}^{NOAH} and ET_{EB}^{ALEXI} is an artifact of erroneously excluding ΔTWS from the ET_{WB} estimate. The results supported the argument that the NOAH land surface model could provide accurate estimates of inter-annual ET variability in humid basins, in contrast to the conclusions of Zhang et al. (2012).

The work in Han et al. (2015) showed the importance of possible errors in ET_{WB} when assessing ET_{EB} estimates. However, their conclusions were limited to the Mississippi River basin and based on 10 years of data, which limits the conclusions that can be made for ET_{EB}^{NOAH} in humid basins across the CONUS. In addition, the study used consistency between ET_{EB}^{NOAH} and ET_{EB}^{ALEXI} as evidence that ET_{EB}^{NOAH} estimates were correct, which ignores the possibility that both products could exhibit similar errors despite their different estimation approaches. We also note that comparisons between ET_{EB}^{NOAH} and spatially interpolated ET estimates from a network of eddy covariance towers (Jung et al. 2009) in the Ohio Basin were poor (correlation coefficient of ~0.4, Han et al. 2015), somewhat countering the claim that discrepancies between ET_{EB}^{NOAH} and ET_{WB} in the Ohio were due to errors in ET_{WB}. Finally, error in P has been shown to be a potentially large source of uncertainty in ET_{WB} as well as in land surface model ET (Hain et al. 2015), yet annual ET_{EB}^{NOAH} was the only ET estimate compared directly to ET_{WB}. Since ET_{EB}^{NOAH} and ET_{WB} used in that study had the same P forcing (from NLDAS-2), basin level correlation between these two ET estimates could have been influenced by correlated errors in P.

In this study, we revisit the results of Han et al. (2015) using an expanded analysis to address some of the limitations of the original analysis. Specifically, we consider the following additions:

1. In addition to using ET_{EB}^{NOAH} and ET_{EB}^{ALEXI}, we include an ET product that uses a similar model formulation to NOAH, but which uses different information on local moisture anomalies and real-time remotely sensed vegetation indices from the MODIS platform to force moisture based reductions in ET (ET_{EB}^{MOD16}, Mu et al. 2011).
2. We compare all 3 ET products (ET_{EB}^{NOAH}, ET_{EB}^{MOD16}, ET_{EB}^{ALEXI}) directly to ET_{WB}, expanding the study area to include 671 basins with substantial coverage over the CONUS.
3. We examine both long-term bias between ET_{WB} and ET_{EB} and inter-annual anomalies between ET_{WB} and ET_{EB} after accounting for long-term bias. By contrasting ET_{EB} performance across products, comparing long term and inter-annual non-closure errors between ET_{EB} and ET_{WB}, and relating these non-closure errors to watershed characteristics, we characterize regional variability in ET_{EB} performance.
4. We integrate a new ΔTWS product derived from the GRACE satellite data (GRACE Mascon, Wiese 2015) to evaluate hydrologic consistency between ET_{EB} and ET_{WB}, and assess the ability to combine ET_{EB} with both GRACE Tellus and Mascon ΔTWS to describe the difference between precipitation and runoff across the CONUS. Regional variability in the importance of ΔTWS for annual water budgets across the CONUS is discussed.
5. To further diagnose the cause of non-closure errors, we assess regional relationships between non-closure errors and other atmospheric and meteorological proxies besides ΔTWS (e.g., precipitation, horizontal heat advection, snow cover) that could be a source of error in either ET_{EB} or ET_{WB} estimates.

2. METHODS

2.1 ET Data Overview

2.1.1 Water balance ET (ET_{WB})—

ET_{WB}. was calculated for the 671 watersheds included in the large-sample basin scale hydrometerological dataset developed by Newman et al. (2015) from the Hydro-Climatic Data Network (HCDN) 1988 dataset (marked “HCDN 2009” in the GAGES-II dataset, Falcone et al. 2010). The selection of these small to mid-sized gaged watersheds (median area 336 km^2) was based on criteria for minimal human disturbance (Falcone et al. 2010) and data quality (including basin-level density of precipitation gages) and continuity. As an additional screening, we validated that basins included in the database had minimal irrigation by calculating the percent irrigated area using the MODIS Irrigated Agriculture Dataset for the United States (MIrAD, Pervez et al. 2010). The vast majority of basins had negligible irrigated area.

Annual water year (October-September) discharge for each gage was calculated between 2003–2015 and converted to annual runoff (Q, mm) by dividing by drainage area. Precipitation data were taken from the National Land Data Assimilation System 2

(NLDAS-2, Xia et al. 2012) primary forcing dataset. Basin-average P was calculated for water years 2003–2015 by averaging all NLDAS grid cells whose centers are located within the basin (Hijmans 2014). For each basin for each year, annual ET_{WB} was calculated using Equation 1.2, as was the 13-year mean ($\overline{ET_{WB}}$). Basins with any ET_{WB} values less than zero were eliminated from the analysis.

2.1.2 Energy Balance Based ET Products (ET_{EB})—In addition to the water budget approach, ET can be characterized using a surface energy budget as follows:

$$Rn = H + LE + G \quad \text{Eq. 2}$$

where Rn (net radiation, generally expressed as W m^{-2}) is partitioned to sensible heat flux (H, W m^{-2}), latent heat flux (LE, W m^{-2}), and ground heat flux (G, W m^{-2}), with the partitioning dependent on surface and atmospheric characteristics and conditions. ET_{EB} (mm) can then be derived from LE by dividing by the latent heat of vaporization of water.

When there is unlimited water available for evapotranspiration, the partitioning of Rn between H, G, and LE can be modeled using a variety of physically-based or empirical methods, where G is generally assumed to be constant (often zero) on a daily or greater time step, and the relative partitioning of energy to H and LE is dependent on surface meteorological conditions, including air temperature, relative humidity, wind speed, and vegetation cover.

While numerous surface energy budget based ET_{EB} exist (see Kalma et al. 2008 for a review), for the purposes of this study, we consider two basic classes:

1. Penman-Monteith based ET (ET_{EB}^{NOAH}, ET_{EB}^{MOD16}): Potential Evapotranspiration (PET) is calculated from daily surface meteorological variables (reanalysis data) using a modified Penman-Monteith equation, with spatially variable vegetation indices used to estimate surface emissivity and resistance. PET is scaled to ET using time varying estimates of moisture availability.
    a. ET_{EB}^{NOAH}(Chen et al., 1996): NLDAS-2 Noah LSM ET output is calculated by the National Centers for Environmental Prediction (NCEP) using North American Regional Reanalysis (NARR) surface meteorological data fields, with NARR downward shortwave radiation bias-corrected to the University of Maryland Surface Radiation Budget (SRB) (Pinker et al. 2003), which uses data from the GOES-8 satellite. Resistance terms are applied to Penman-based PET for four separate terrestrial surface types (canopy transpiration, canopy evaporation of intercepted precipitation, soil evaporation, sublimation). Resistance terms are modified by the underlying soil moisture, which is forced by NLDAS-2 P (Xia et al. 2012, Xia et al. 2015, Peters-Lidard et al. 2011).
    b. ET_{EB}^{MOD16}(Mu et al. 2011): The ET_{EB}^{MOD16} product developed by the University of Montana uses Global Modeling and Assimilation Office (GMAO) Modern-Era Retrospective analysis for Research and

Applications (MERRA, Rienecker et al., 2011) data, in conjunction with 8-day data inputs from the MODIS platform of the AVHRR satellite (leaf area index, enhanced vegetation index, and albedo) to calculate PET. As vegetation growth is primarily limited by water availability, real-time updates of local vegetation conditions from MODIS data are used to capture water limited ET (Cleugh et al. 2007). Mu et al. (2007) updated this algorithm to scale evaporation by relative humidity and scale transpiration by vapor pressure deficit, used as a proxy indicator of soil moisture limitations

2. Remotely-sensed thermal infra-red ET (ET_{EB}^{ALEXI}, Anderson et al. 1997, Kustas et al., 2011): H is calculated from a time-differential linear model of atmospheric boundary layer (ABL) development which is forced with two measurements of radiometric surface temperature taken in the morning from GOES satellites, an initial estimate of ABL height from early morning sounding data, and wind speed to give an estimate of instantaneous H flux at 1.5 hour prior to solar noon. Instantaneous LE is then calculated using Equation 2, and is converted to daily cumulative LE using ancillary estimates of daily cumulative insolation assuming a diurnally constant value of (LE/Sdn), where Sdn is instantaneous solar radiation.

2.2 Long-term Non-Closure Error

To examine long-term non-closure errors, the period of record bias was calculated as:

$$\overline{\varepsilon_{NC}} = \overline{ET_{EB}} - \overline{ET_{WB}} \quad \text{Eq.3}$$

where the overbar indicates the 2003–2015 mean, and ET_{EB} is based on either NOAH, MOD16, or ALEXI estimates. We first evaluate and compare the distribution and spatial patterns of $\overline{\varepsilon_{NC}}$ across the three energy-balance models.

We then calculate the rank correlation between $\overline{\varepsilon_{NC}}$ and a variety of watershed characteristics to determine the potential drivers of long-term bias. The majority of these watershed characteristics were extracted from the Geospatial Attributes of Gages for Evaluating Streamflow, version II (GAGES II) database (Falcone et al 2010), with several exceptions. For example, percent (%) irrigated area in each basin, a potential source of anthropogenic decoupling of true ET from ET_{WB}, was taken as the basin mean value for the MIrAD-US dataset (Pervez and Brown 2010). Mean horizontal heat advection ($\overline{H_{HORIZ}}$, a possible error source in ET_{EB}) was calculated as the basin mean of the square root of the sum of squared northward and eastward components of vertically integrated sensible heat flux for June, July, and August (JJA) for the years 2003–2015 from the ERA-Interim reanalysis product (Dee et al., 2011). For potential long-term water balance errors, mean ΔTWS for each basin was calculated using the method in Landerer and Swenson (2012) from 2003–2015 for the GRACE Tellus (Swenson, 2012) and the newer GRACE Mascon (Save et al. 2016) data products (see Supplemental Text 1 for more detail). Since both ET_{EB}^{MOD16} and ET_{EB}^{ALEXI} depend on remotely sensed data that may not be available on certain

days (e.g., because of cloud cover), estimates depend on gap-filling procedures that could degrade their accuracy. Therefore, we used the average percent of pixel-days with missing values per basin per year (mean % NA pixel-days) from the direct retrievals from GOES thermal imagery in the ALEXI product as a proxy for obstructive cloud cover.

Finally, to ensure that the results are not sensitive to the NLDAS-2 precipitation product used in the calculation of $\overline{ET_{WB}}$, we reevaluate the distribution of $\overline{\varepsilon_{NC}}$ with $\overline{ET_{WB}}$ calculated from several different precipitation datasets (see Supplemental Text 2).

2.3 Inter-annual Non-Closure Error

Inter-annual non-closure errors can arise from a variety of error sources in ET_{EB} or ET_{WB}. In the case that these errors are caused by bias in the mean or variance of ET_{EB} or $ET_{WB,}$ it may still be possible to show hydrologic consistency in the direction of inter-annual change between ET_{EB} and ET_{WB} if the data are first mean adjusted and scaled. Therefore, we center both variables to remove long-term mean biases (with prime indicating annual departure from the mean, $ET' = ET - \overline{ET}$), and also produce a scaled version of these data by dividing by the standard deviation (denoted $sET' = \frac{(ET - \overline{ET})}{sd(ET)}$). We then use geographically weighted regression (GWR) to explore how the relationship between ET_{WB} and ET_{EB} varies across space. GWR is a localized regression method that uses a kernel function and spatial bandwidth to pool data at nearby sites in the estimation of a regression linking response and predictor variables (Brunsdon et al. 1998). The resulting regression exhibits smooth variations in the coefficient estimates at different sites across space. GWR provides a compromise between calculating a single regression coefficient on the global dataset, which would eliminate spatial variability in the process, or calculating individual regression estimates for each point in the study area, which could result in unstable coefficient estimates with high standard errors due to low sample size. In this analysis, GWR models were built to explore the spatial variability in how well ET'_{EB} tracks inter-annual variability in ET'_{WB}.

For each ET_{EB} product, two GWR models were built:

$$sET'_{EB,i,t} = \beta_i sET'_{WB,i,t} + \varepsilon_{i,t} \quad \text{(Eq. 3.1)}$$

$$ET'_{EB,i,t} = \beta_i ET'_{WB,i,t} + \varepsilon_{i,t} \quad \text{(Eq. 3.2)}$$

where t indicates the year and i indicates the i^{th} basin. β_i is allowed to smoothly vary across the CONUS by pooling nearby basins using a Gaussian kernel function with optimal bandwidth selected via cross validation (Bivand and Yu, 2017). The first model (Eq. 3.1) explores whether sET'_{EB} exhibits hydrologic consistency with sET'_{WB} after accounting for the potential of both mean and variance biases in all water and energy balance terms The second model (Eq. 3.2) explores the ability of ET'_{EB} to close inter-annual terrestrial water budgets in an absolute sense (less any mean biases in either ET_{EB} or ET_{WB}). In particular, β_i from Eq. 3.2 not only reflects the strength of the relationship between ET'_{EB} and ET'_{WB}, but also the

difference in magnitude of inter-annual fluctuations between energy-balance and water-balance ET (i.e., variance bias).

To better understand the distribution of errors exhibited by the GWR, we also examine the distribution of scaled, inter-annual non-closure errors, calculated as the difference between the scaled, annual anomalies:

$$ {}_s\varepsilon_{\mathrm{NC}} = s\mathrm{ET}'_{EB} - s\mathrm{ET}'_{WB} \quad \text{(Eq. 4)} $$

For a particular region, any clustering in ${}_s\varepsilon_{\mathrm{NC}}$ in certain years and not others would suggest that characteristics of specific climate events contribute substantially to discrepancies between $s\mathrm{ET}'_{EB}$ and $s\mathrm{ET}'_{WB}$. These discrepancies could arise from errors in both energy-balance and water-balance based models. Likely sources of error in annual estimates of $\mathrm{ET_{WB}}$ include errors in estimates of basin P or violations in the assumption that changes in basin surface water storage (ΔTWS) are negligible on an annual time-step. Energy balance models were not originally designed to operate at the continental scale, where horizontal fluxes of energy can lead to violations of the assumption of daily local conservation of energy (Trenberth et al. 2009). In addition, these models have parameterizations of sublimation that have been previously questioned (Wang et al. 2015). Therefore, we explore clustering in ${}_s\varepsilon_{\mathrm{NC}}$ by product and year, and then correlate ${}_s\varepsilon_{\mathrm{NC}}$ to several variables on an annual scale, including basin-average P, ΔTWS for both GRACE Tellus ($\Delta\mathrm{TWS_T}$) and GRACE Mason ($\Delta\mathrm{TWS_M}$) data products, % of precipitation as snow from the ERA-Interim monthly data, and JJA horizontally advected sensible heat flux ($\mathrm{H_{HORIZ}}$).

3. RESULTS

3.1 Long-term non-closure error

The mean annual estimates for all water and energy balance models ($\overline{\mathrm{ET_{WB}}}$, $\overline{\mathrm{ET_{EB}^{ALEXI}}}$, $\overline{\mathrm{ET_{EB}^{NOAH}}}$, and $\overline{\mathrm{ET_{EB}^{MOD16}}}$) show similar spatial patterns, with the highest and lowest mean annual ET estimates seen in the Southeast and Southwest and western continental interior, respectively (Figure 1a). However, $\mathrm{ET_{EB}}$ models exhibit several areas of disagreement in estimates of long-term mean annual ET with respect to $\mathrm{ET_{WB}}$, as well as with each other (Figure 1b). For example, in basins to the west of the Rockies along the California coast, $\overline{\mathrm{ET_{EB}^{ALEXI}}}$ predicts higher ET values compared to $\overline{\mathrm{ET_{EB}^{MOD16}}}$ and $\overline{\mathrm{ET_{EB}^{NOAH}}}$, which are similar to those seen in the $\mathrm{ET_{WB}}$ model. Over the entire CONUS, $\overline{\mathrm{ET_{EB}^{NOAH}}}$ tends to predict lower annual ET compared to the other three models. $\overline{\mathrm{ET_{EB}^{MOD16}}}$ and $\overline{\mathrm{ET_{EB}^{ALEXI}}}$ tend to be higher than $\overline{\mathrm{ET_{WB}}}$ in the Pacific Northwest and in basins along the Appalachian Mountain Chain, while $\overline{\mathrm{ET_{EB}^{ALEXI}}}$ is higher than all other ET estimates along the Gulf Coast and through Florida, and $\overline{\mathrm{ET_{EB}^{MOD16}}}$ is lower than all other models in the central U.S. The range (maximum-minimum) of ET estimates across the CONUS is smaller for $\overline{\mathrm{ET_{EB}^{MOD16}}}$ (924 mm) and $\overline{\mathrm{ET_{EB}^{NOAH}}}$ (881mm) compared to $\overline{\mathrm{ET_{EB}^{ALEXI}}}$ (1060 mm), which is more similar to that of $\overline{\mathrm{ET_{WB}}}$ (1136

mm). In general, spatial consistency between the long-term means established by the four ET models appears to be lowest in the eastern third of the country and along the western coastline.

When analyzing Figure 1, we note that errors in long-term basin-average P (and thus $\overline{ET_{WB}}$) are expected for any particular basin, for example due to gage location within basin topography. It is less likely, however, that long-term error in orographically-adjusted basin-averaged P would be biased in the same direction across the 671 basins, although this possibility cannot be precluded due to documented undercatch of gage precipitation linked to surface winds (Adam and Lettenmaier 2003, Yang et al. 2005) and/or potential systematic errors in data processing. However, the latter issue can be controlled for through the comparison of multiple products in the calculation of $\overline{ET_{WB}}$. Further, if the mean of the distribution of $\overline{\varepsilon_{NC}}$ is significantly above zero for some ET_{EB} products and below zero for others, we argue that one or more of these ET_{EB} products are systematically biased compared to the true (and unknown) long-term ET.

In this context, we examine the distribution of $\overline{\varepsilon_{NC}}$ for each of the three energy balance models in Figure 1c. The mean of $\overline{ET_{EB}^{MOD16}}$ is relatively unbiased relative to $\overline{ET_{WB}}$, with a mean bias of about 4% above $\overline{ET_{WB}}$. The mean ALEXI-based ET estimates are approximately 120 mm/year higher than mean $\overline{ET_{WB}}$ on average, suggesting that $\overline{ET_{EB}^{ALEXI}}$ has a positive mean bias of 21% across CONUS. However the majority of the basins with the highest values of $\overline{\varepsilon_{NC}^{ALEXI}}$ are in mountainous regions (i.e. along the Cascades, Sierra Nevada, Rockies, and Appalachian Mountains), where topographic (slope and aspect) effects on the assumed net radiation (Rn in Eq. 2) will impact the information content of remotely sensed data inputs. Removing basins with average slopes above 15% from the analysis (Supplemental Figure S1) reduces the positive mean bias of $\overline{ET_{EB}^{ALEXI}}$ to approximately 49 mm/year (or 9%) above mean $\overline{ET_{WB}}$.

While the ALEXI and MOD16 products produce slightly higher long-term ET estimates on average than the water balance based estimate, which is potentially consistent with undercatch in the NLDAS-2 precipitation product, the mean $\overline{ET_{EB}^{NOAH}}$ is about 80 mm/year lower than mean $\overline{ET_{WB}}$ on average (negative mean bias of about 13%). Importantly, since both $\overline{ET_{WB}}$ and $\overline{ET_{EB}^{NOAH}}$ are based on the same NLDAS-2 precipitation dataset, it is unlikely that this discrepancy is linked to systematic bias in the precipitation field. Further, the results are very similar if the scatterplots in Figure 1c are replicated for $\overline{ET_{WB}}$ calculated with three additional gridded precipitation products (Supplemental Text 2, Figure S2). A notable exception occurs when using the less accurate NLDAS-2 secondary precipitation forcing, which is based on the NARR reanalysis and is thus likely to be less accurate than gaged data. In this case, the NOAH model exhibits the least long-term bias compared to ET_{WB}. Assuming that any systematic, long-term bias in $\overline{ET_{WB}}$ across the majority of basins would be a downward bias associated with undercatch, the results in Figure 1 indicate that ALEXI and MOD16 models are more likely to be relatively unbiased with respect to the true, long-

term ET across the CONUS, particularly in regions of mild terrain. Under this same assumption, Figure 1 also suggests that there may be moderate negative biases in the mean long-term annual ET estimated by the NOAH model.

The assessment above does not consider ΔTWS as a potential source of error in ET_{WB} but this error should be minimized in the long-term average $\overline{ET_{WB}}$ because it is unlikely that there would be large changes in basin storage in basins with relatively little anthropogenic influences over a decadal period (i.e., $\overline{\Delta TWS}$ should be near zero). To validate this assumption, 2003–2015 $\overline{\Delta TWS}$ was calculated using the GRACE Tellus $(\overline{\Delta TWS_T})$ and GRACE Mascon $(\overline{\Delta TWS_M})$ products (Figure 2, left column). We found that for most regions in the CONUS, $\overline{\Delta TWS}$ was very close to zero, with an interquartile range across basins of −1.5 – 6.4 mm for $\overline{\Delta TWS_T}$ and 0.1–11.8 mm for $\overline{\Delta TWS_M}$. It should be noted, however, that the standard deviations of $\overline{\Delta TWS}$ (Figure 2, right column) suggests that inter-annual variability in ΔTWS may be important in some regions, and estimates are higher for $\overline{\Delta TWS_M}$ than for $\overline{\Delta TWS_T}$. $\overline{\Delta TWS_M}$ indicates that the highly irrigated regions of southern California and near the Ogallala Aquifer are losing between approximately 20–30mm of TWS per water year, and that TWS in the Midwest may be increasing by about 10–20 mm per year. However, large $\overline{\varepsilon_{NC}}$ values for all of the energy balance products are found in regions where the absolute magnitudes of $\overline{\Delta TWS}$ are negligible, and even in regions with high-magnitude $\overline{\Delta TWS}$, these values are still several times smaller than $\overline{\varepsilon_{NC}}$ (Figure S3). Further, changes in basin storage linked to irrigated water withdrawn from hydraulically disconnected groundwater sources (e.g., Ogallala Aquifer) should not substantially affect the ET in relatively natural basins without much irrigation. Therefore, it does not seem likely that $\overline{\Delta TWS}$ plays a large role in the long-term biases between ET_{WB} and ET_{EB} estimates.

To identify regional characteristics that could explain long-term non-closure error ($\overline{\varepsilon_{NC}}$), we evaluated the Spearman correlation between $\overline{\varepsilon_{NC}}$ and physical and meteorological characteristics associated with each watershed (Figure 3). We note that absolute correlation coefficients above 0.08 are statistically significant ($\alpha = 0.05$) in a sample of 671 basins, but we focus on stronger relationships with absolute correlation coefficients above 0.4–0.6 that indicate potentially more important relationships between watershed characteristics and $\overline{\varepsilon_{NC}}$. Since many watershed characteristics are collinear, we interpret large correlations between $\overline{\varepsilon_{NC}}$ and these characteristics as indicative of a strong association, but not necessarily a causal relationship.

One of the most notable results of the correlation analysis is that $\overline{\varepsilon_{NC}^{ALEXI}}$ and $\overline{\varepsilon_{NC}^{MOD16}}$ are strongly and positively correlated with mean annual P. In contrast, $\overline{\varepsilon_{NC}^{NOAH}}$, which is forced with the same NLDAS-2 P used to calculate $\overline{ET_{WB}}$, shows no significant correlation with mean annual P. This result suggests that long-term bias in NLDAS-2 P (and thus $\overline{ET_{WB}}$) for each basin may be a leading driver of long-term non-closure error for the ALEXI and MOD16 models, but not for the NOAH model because it is based on the same precipitation

dataset. This conclusion is consistent with the smaller variance in the scatter of Figure 1c for the NOAH model compared to the ALEXI and MOD16 models.

Another notable result is that both $\overline{\varepsilon_{NC}^{ALEXI}}$ and $\overline{\varepsilon_{NC}^{MOD16}}$ show significant positive correlation with increasing mean percent slope in the basin, a result that may indicate degradation of remotely sensed information in basins with complex terrain (also see Figure S1). Alternatively, correlation between $\overline{\varepsilon_{NC}^{ALEXI}}$ and $\overline{\varepsilon_{NC}^{MOD16}}$ and % SLOPE may be further evidence that NLDAS-2 P, and therefore $\overline{ET_{WB}}$, are negatively biased in regions with complex terrain. Though cloud cover has been associated with degraded performance of $\overline{ET_{EB}^{ALEXI}}$ at shorter time scales, mean % NA pixel-days over each basin was not significantly correlated to $\overline{\varepsilon_{NC}^{ALEXI}}$, suggesting that cloudiness does not contribute to long-term model bias. However, $\overline{\varepsilon_{NC}^{MOD16}}$ was significantly and negatively correlated to mean % NA pixel-days, suggesting that $\overline{ET_{EB}^{MOD16}}$ may be more likely to underestimate ET in cloudier basins. This discrepancy may be explained by the different ways that the ALEXI and MOD16 models gap-fill remotely sensed data.

Many of the remaining correlation coefficients are small, suggesting these relationships have little biophysical importance, though some are interesting to observe in the context of contrasting model formulation. For example, mean JJA horizontal heat advection ($\overline{H_{HORIZ}}$) showed significant correlation with $\overline{\varepsilon_{NC}^{MOD16}}$, but not $\overline{\varepsilon_{NC}^{ALEXI}}$. This is interesting, as an increase in $\overline{H_{HORIZ}}$ (i.e., large-scale heat advection) would be associated with an increase in estimation of LE and ET in the MOD16 model, whereas increasing $\overline{H_{HORIZ}}$ would be associated with a direct decrease in estimates of LE and ET (and an increase in estimates of H) in the ALEXI model.. Also, weak correlations between $\overline{\varepsilon_{NC}}$ and $\overline{\Delta TWS_M}$ and $\overline{\Delta TWS_T}$, particularly for $\overline{\varepsilon_{NC}^{ALEXI}}$ and $\overline{\varepsilon_{NC}^{NOAH}}$, further suggest that changes in basin storage are not driving discrepancies between water-budget and energy-budget based estimates of average annual ET (also see Figure S3)

3.2 Inter-annual non-closure error

We first evaluate the relationship between annual ET_{EB} and ET_{WB} estimates for water years 2003–2015 over the CONUS using GWRs. By contrasting the β estimates for NOAH, ALEXI, and MOD16 products and looking for regional coherence and discordance in these estimates, we hope to elucidate regions where hydrologic inconsistency is caused by annual errors in either ET_{EB} or ET_{WB}. In particular, hydrologic consistency between ET_{WB} and ET_{EB}^{NOAH} would be expected to be higher than hydrologic consistency between ET_{WB} and both ET_{EB}^{ALEXI} or ET_{EB}^{MOD16} in basins where annual error in basin-level estimates from NLDAS-2 P were of sufficient magnitude to lead to correlated errors in ET_{EB}^{NOAH} and ET_{WB} estimates.

Figure 4 shows the spatial distribution of β (4a) and R^2 (4b) for the GWR on scaled data (i.e., sET'_{WB} against sET'_{EB}, Equation 3.1). These spatial patterns are validated with location-specific regressions and Pearson correlation coefficients (see Supplemental Figure S4) to ensure that results are not an artifact of the spatial bandwidth chosen in the GWR models (Wheeler and Tiefelsdorf, 2005). Several patterns emerge when looking at the spatial distributions in Figure 4 a,b. First, the eastern portions of the Missouri, Arkansas White-Red, and Texas-Gulf River basins have β values that approach one, suggesting that for all three products, ET_{WB} and ET_{EB} track each other very closely. For all three models, β values are also moderately high for basins in the Lower Mississippi River, Lower Colorado, and the southern part of the California basin. However, in all other regions of the CONUS, the variability in ET_{EB} relative to ET_{WB} diverges, and this divergence varies by energy balance model. For all ET_{EB} estimates, β values are lower in the east than in the continental interior and southwest coast. In addition, β and R^2 values are higher for ET_{EB}^{NOAH} than for ET_{EB}^{ALEXI} and ET_{EB}^{MOD16} in the Pacific Northwest and across most regions east of the Mississippi. For instance, where β values are below 0.1 for ET_{EB}^{MOD16} and ET_{EB}^{ALEXI} for several basins in the New England and Mid-Atlantic regions (suggesting no hydrologic consistency), they range between 0.1 and 0.4 for ET_{EB}^{NOAH}. This suggests that in these regions, inter-annual errors in NLDAS-based basin P likely play a role in the variability of non-closure errors. Still, the β (R^2) values for the NOAH model are consistently below 0.5 (0.25) for many eastern regions of the CONUS, indicating that there are other drivers of non-closure error besides errors in P.

We also explore how the variance differs between ET_{EB} and ET_{WB}. Pooling data from all gages across the CONUS, the inter-annual variance of ET_{WB} is (17167.8), which is much larger than that for MOD16 (1344.1), NOAH (1890.9), and to a lesser extent, ALEXI (2751.6). The catchment level standard deviations of all ET estimates, as well as the ratios of variances between ET_{EB} and ET_{WB} for each basin are shown in Supplemental Figure S5. This variance bias is also reflected in Figure 4c and d, which shows the β and R^2 values for the GWR on unscaled data (Eq 3.2). These β values are substantially lower than those for the scaled data (Figure 4a), highlighting the large variance biases between ET_{WB} and ET_{EB}. Figure 4 and Figure S5 show that, in regions where hydrologic consistency between sET'_{WB} and sET'_{EB} is highest (see Figure 4a), the variance of ET_{EB} is about a quarter of that for ET_{WB}, regardless of energy balance model. In regions where hydrologic consistency is moderate or low, the variance of ET_{EB} is almost two orders of magnitude less than that of ET_{WB}. This suggests that there is significant variance bias in either or both of the energy and water balance based estimates. We speculate that for the water balance model, this may be related to inflated variance in basin-averaged estimates of annual precipitation.

Figure 4 and the comparison of β and R^2 values between the energy balance models indicate that annual errors in basin P have a large role in non-closure errors, but other factors are also likely important. To explore this further, we first examine the distribution $_s\varepsilon_{NC}$ in individual years to determine if there are spatial patterns in the error term. Any clustering in $_s\varepsilon_{NC}$ in certain regions and years would suggest that characteristics of specific climate events contribute substantially to discrepancies between sET'_{EB} and sET'_{WB}. Figure 5 shows the

values of ${}_s\varepsilon_{NC}$ for 2 years with large-magnitude errors (2006 and 2012), with all years shown in Figure S6. Clustering in ${}_s\varepsilon_{NC}$ is clearly apparent. For instance, in 2012, all three sET'_{EB} estimates were well above sET'_{WB} in the Northeast region, whereas in 2006 the opposite trend occurs. Further, in 2006, the errors are of much higher magnitude for MOD16 than the other products. Clustering of higher magnitude annual ${}_s\varepsilon_{NC}$ appears to be strongest in regions where hydrologic consistency between ET_{EB} and ET_{WB} is lowest (Figure 4).

To further diagnose this clustering and the cause of inconsistencies (i.e., low β and R^2 values and low variance ratios) revealed in the regressions above, scaled, non-closure errors (${}_s\varepsilon_{NC}$) are related to a series of atmospheric and meteorological proxies that could represent the cause of errors in ET_{WB} or ET_{EB} estimates. The first row of Figure 6 shows the correlation between ${}_s\varepsilon_{NC}$ and basin P. ${}_s\varepsilon_{NC}^{ALEXI}$ shows broad negative correlations with basin P estimates across CONUS, particularly in the Northeast, suggesting that scaled anomalies of ET_{EB}^{ALEXI} are low relative to scaled anomalies in ET_{WB} when P is above average. For ${}_s\varepsilon_{NC}^{MOD16}$ and ${}_s\varepsilon_{NC}^{NOAH}$, the pattern of correlation coefficients for P is similar to that for ALEXI but somewhat muted, particularly along the Rocky Mountains. These results could be interpreted to suggest that when estimates of P are either too high or too low, the error propagates into ET_{WB} and contributes to non-closure errors (i.e., we have high variance bias in P). Alternatively, it might suggest that errors in ET_{EB} could be anti-correlated with P.

However, the similarity of the relationship between P and both ${}_s\varepsilon_{NC}^{MOD16}$ and ${}_s\varepsilon_{NC}^{NOAH}$ is somewhat surprising, given similar formulations of the NOAH and MOD16 products and that the NOAH product is the only ET_{EB} model based on the same P data as the ET_{WB} model. This may be partially explained by correlations between the precipitation products used by NOAH (NLDAS-2) and MOD16 (GMAO). This may also be partially explained when examining the correlations between ΔTWS and ${}_s\varepsilon_{NC}$. Like Han et al. (2015), we examine how ${}_s\varepsilon_{NC}$ relates to ΔTWS, but examine this relationship for both the Tellus and Mascon GRACE products (Figure 6, rows 2 and 3) to evaluate whether recent advances in processing of GRACE data has improved ΔTWS estimates. The ${}_s\varepsilon_{NC}$ values for all three models show high negative correlation with both GRACE products in the Souris Red Rainy, Great Lakes, New England, and Mid-Atlantic regions. ${}_s\varepsilon_{NC}^{NOAH}$ and ${}_s\varepsilon_{NC}^{ALEXI}$ also show strong, negative correlations with ΔTWS_M and ΔTWS_T in the Upper Mississippi basin, while ${}_s\varepsilon_{NC}^{MOD16}$ and ${}_s\varepsilon_{NC}^{ALEXI}$ show strong negative correlations with both GRACE products in the Pacific Northwest. The results are very similar across both GRACE products, although some differences can be seen (e.g.,for MOD16 in the Ohio River basin), and the relationships are slightly stronger with the newer Mascon data compared to the Tellus data (see Figure S7), which exhibits slightly more inter-annual variability than Tellus (Figures 2 and S7). We note these results were seen in spite of the fact that many basins used in this study were well below the minimum spatial resolution of the GRACE products. We also note that if ΔTWS is included directly in the ET_{WB} estimates, the sign of ${}_s\varepsilon_{NC}$ generally does not change, and there are only modest impacts to the magnitude of non-closure errors (Figures S8 and S9). These modest impacts may be caused by a variance bias in either P or ΔTWS, motivating the use of the correlation analysis in Figure 6 to control for these potential biases and maximize the potential signal between non-closure errors and ΔTWS (Supplemental Text 3).

Overall, the correlations between ${}_s\varepsilon_{NC}$ and ΔTWS are strongest in many of the same basins with strong correlations between P and ${}_s\varepsilon_{NC}$. These results indicate that the correlations between P and ${}_s\varepsilon_{NC}$ may be the result of errors in ET_{WB} linked to ΔTWS that are strongest when conditions are wet. This makes it difficult to assess whether errors in P or the omission of ΔTWS is driving the non-closure error. Furthermore, both NOAH and MOD16 scale ET by moisture availability, and so in theory, errors in this scaling process could lead to the correlations between ${}_s\varepsilon_{NC}$, P, and ΔTWS seen in Figure 6. However, ALEXI does not scale ET based on moisture and still has similar correlations between ${}_s\varepsilon_{NC}$, P, and ΔTWS. Therefore, the strong relationships between ${}_s\varepsilon_{NC}$ and both P and ΔTWS suggest that some combination of uncertainties linked to ET_{WB} play a large role in non-closure errors.

We are also interested in whether violations of assumptions in the energy balance models drive non-closure errors, particularly as they relate to regional climate phenomena that could translate to regional clustering of ${}_s\varepsilon_{NC}$. We consider two such phenomena: % P from snow (Figure 6, row 4) and large-scale horizontal heat advection (Figure 6, row 5). ${}_s\varepsilon_{NC}$ shows positive correlations to % P as snow in the Ohio River and northern Mid-Atlantic regions for all models, suggesting that in these areas, overestimation of ${}_sET'_{EB}$ relative to ${}_sET'_{WB}$ was higher in years when more P came as snowfall. This relationship is strongest for MOD16, suggesting that wintertime ET (associated with sublimation) may be overestimated in this routine. It should be noted, however, that P undercatch is accentuated during snowy conditions, which could further bias ET_{WB} downwards (Yang et al. 2005, Pan et al. 2003).

An interesting pattern also emerges when ${}_s\varepsilon_{NC}$ is compared to H_{HORIZ}. Years when H_{HORIZ} is high seem to be associated with underestimation of all three sET'_{EB} relative to sET'_{WB} in the Mid-Atlantic, southern Ohio, Tennessee, and South Atlantic Gulf regions. In contrast, large values of H_{HORIZ} appear to be associated with an overestimation of sET_{EB}^{MOD16} relative to sET_{WB} at the corner of the Missouri, Upper Mississippi, Arkansas White-Red, and Lower Mississippi regions. Interestingly, in these same regions, MOD16 appears to differ most strongly from other ET products in general, though more research is needed to identify the reasons for this discrepancy. Overall, errors in P and ΔTWS appear more related to non-closure errors than % P as snow and H_{HORIZ} for most products and regions, although all variables appear important in at least some regions.

4. DISCUSSION AND CONCLUSION

In this study, long-term average and inter-annual ET from three gridded energy balance products were evaluated against ET calculated from a simple terrestrial water budget in 671 basins across CONUS for water years 2003–2015. For long-term biases, $\overline{ET_{EB}^{ALEXI}}$ showed evidence of some overestimation of annual ET relative to $\overline{ET_{WB}}$, although this bias was mostly limited to basins with complex terrain. $\overline{ET_{EB}^{NOAH}}$ showed evidence of some systemic underestimation of annual ET across the entire CONUS region, while the mean of $\overline{ET_{EB}^{MOD16}}$ was very close to the mean of $\overline{ET_{WB}}$. All three $\overline{ET_{EB}}$ products seem capable of tracking ET climatology in the central part of the country, albeit with some underestimation from MOD16, while the three $\overline{ET_{EB}}$ products differed from $\overline{ET_{WB}}$ and each other in the eastern

third of CONUS and along the west coast, a result consistent with previous analysis (Anderson et al. 2013). Our examination of rank correlations between $\overline{\varepsilon_{NC}}$ and watershed characteristics (Figure 3) suggested that 1) $\overline{ET_{EB}^{ALEXI}}$ is overestimated in basins with high mean annual precipitation and complex terrain, 2) $\overline{ET_{EB}^{MOD16}}$ is also overestimated in basins with higher mean precipitation and complex terrain, but also underestimates ET in basins with more cloud cover (% NA retrievals from ALEXI as a proxy), and overestimates ET in basins with more summertime vertically integrated heat flux ($\overline{H_{HORIZ}}$).

Importantly, the high correlation coefficients seen between mean annual precipitation and $\overline{\varepsilon_{NC}^{ALEXI}}$ and $\overline{\varepsilon_{NC}^{MOD16}}$ where absent for $\overline{\varepsilon_{NC}^{NOAH}}$. Since NOAH has NLDAS-2 P as forcing, and is of similar model formulation to MOD16, the differences between these two products may highlight regions where long-term errors in P play an important role in depressing the bias term $\overline{\varepsilon_{NC}^{NOAH}}$. Accordingly, the regions with most prominent contrasts in bias terms between $\overline{ET_{EB}^{NOAH}}$ and $\overline{ET_{EB}^{MOD16}}$ (as well as $\overline{ET_{EB}^{ALEXI}}$) are in mountainous regions of the Eastern US and Pacific Northwest, where gage placement is expected to have an outsized role in long-term bias in gridded P estimates. However, information content of the remotely sensed inputs to MOD16 and ALEXI may be degraded in regions of complex terrain, so it is difficult to distinguish if long-term non-closure error in complex terrain is related to degradation of remotely sense data inputs or error in precipitation

This analysis also confirms that at very long time scales (10+ years, Figure 2), $\overline{\Delta TWS}$ is a negligible part of the terrestrial water budget in most of the basins considered, reducing a potential error term in $\overline{ET_{WB}}$. Taken together, the results above suggest that, at least on a long-term basis, the accuracy of at least some of the $\overline{ET_{EB}}$ estimates are degraded, particularly in humid regions. This is consistent with previous analyses of long-term bias (Sheffield et al. 2009, Gao et al. 2009). Further, we argue that the analysis presented here points most strongly to some degradation for the NOAH product, given that it shares the same precipitation forcing as and is biased downward against $\overline{ET_{WB}}$--a bias which cannot be explained by known, systematic sources of precipitation error (e.g., undercatch).

Many studies have indicated that ET_{EB} cannot be used to close annual terrestrial water budgets (Zhang et al. 2012, Zeng et al. 2014, Velpurri et al. 2013, Han et al. 2015, Liu et al. 2016). Our analysis indicates that in regions with water-limited ET and low inter-annual ET_{WB} variability—specifically in the central US ranging from Texas up through the High Plains—all three ET_{EB} products demonstrated hydrologic consistency with ET_{WB}. Even in these areas, however, the ET_{EB} products were still only able to capture less than half the magnitude of scaled inter-annual variance estimated by ET_{WB} Also, in many regions of the country like the Eastern US and the Pacific Northwest, no or weak consistency is seen between sET'_{EB} and sET'_{WB}. In addition, the variance of unscaled ET_{WB} is twice to more than ten times the variance of unscaled ET_{EB} estimates throughout the CONUS, although this variance bias is smallest with the ALEXI product.

Clustering was seen in basin-level ${}_s\varepsilon_{NC}$ in certain regions and years, a phenomenon that suggests some of these errors are linked to specific climate events that contribute substantially to discrepancies between ET_{EB} and $ET_{WB.}$ To explore whether errors in ET_{WB} or ET_{EB} were the cause of clustering in non-closure errors, we correlated ${}_s\varepsilon_{NC}$ with four key variables that might drive these errors in either ET_{WB} (P, ΔTWS), or ET_{EB} (% P as snow, and H_{HORIZ}). Results showed interesting patterns of correlation between ${}_s\varepsilon_{NC}$ and all of these variables, although correlations were largest and most widespread with P and ΔTWS. This was despite the mismatch in resolution between the GRACE-based ΔTWS estimates and basin sizes. These results generally support the conclusions of Han et al., 2015 that uncertainties in water budget ET estimates play a large role in inter-annual non-closure errors, although these conclusions should likely be modulated somewhat given the long-term biases in energy budget based methods and potential sources of inter-annual errors linked to sublimation and large-scale heat advection.

Further research is needed to identify whether large-scale climatic drivers of terrestrial ET, including phenomena that contribute to atmospheric coupling and the thresholds for energy limited ET, are adequately parameterized in continental-scale energy balance models. A significant question that remains is whether the variance bias contributing to ε_{NC} across CONUS is due to overestimation of inter-annual variability in ET_{WB} or underestimation of inter-annual variability in ET_{EB}. While some of this bias is due to inter-annual fluctuation in basin ΔTWS and inflated variance bias in estimates of basin P, it is unclear if this is the sole source. As ET is commonly assumed to have relatively low variability on an inter-annual basis, resolution to this question could serve to validate or challenge a critical assumption of terrestrial hydrologic models.

The modestly improved relationships seen between non-closure errors and the higher resolution GRACE Mascon data over the Tellus data suggest that further improvements in the resolution of ΔTWS estimates may be needed to help resolve these questions. We feel the mismatch in resolution between the ΔTWS data and basin sizes used in this study was a major limitation of the approach, despite the significant signals that were uncovered. Therefore, as new and higher resolution GRACE products continue to be developed, we argue for the importance of revisiting ET_{WB} and ET_{EB} comparisons in future work using this data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

NLDAS-2 data used in this study, including data from the NOAH LSM, were acquired as part of the mission of NASA's Earth Science Division and archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). GRACE land are available at http://grace.jpl.nasa.gov, supported by the NASA MEaSUREs Program. MOD16 ET data were acquired from the Numerical Terradynamical Simulation Group (NTSG) at the University of Montana with availability supported by NASA.

6. REFERENCES

Adam JC, & Lettenmaier DP (2003). Adjustment of global gridded precipitation for systematic bias. Journal of Geophysical Research: Atmospheres, 108(D9).

Anderson MC, Norman JM, Diak GR, Kustas WP, & Mecikalski JR (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 MC, Kustas WP, Alfieri JG, Gao F, Hain C, Prueger JH, … & Chávez JL (2012). Mapping daily evapotranspiration at Landsat spatial scales during the BEAREX'08 field campaign. Advances in Water Resources, 50, 162–177.

Anderson MC, Hain C, Otkin J, Zhan X, Mo K, Svoboda M, … & Pimstein A (2013). An intercomparison of drought indicators based on thermal remote sensing and NLDAS-2 simulations with US Drought Monitor classifications. Journal of Hydrometeorology, 14(4), 1035–1056.

Bivand R, & Yu D (2017). spgwr: Geographically Weighted Regression. R package version 0.6–31/r1743.

Brunsdon C, Fotheringham S, & Charlton M (1998). Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician), 47(3), 431–443.

Cai X, Yang ZL, David CH, Niu GY, & Rodell M (2014). Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin. Journal of Geophysical Research: Atmospheres, 119(1), 23–38.

Chen F, Mitchell K, Schaake J, Xue Y, Pan H, Koren V, Duan Y, Ek M, and Betts A, 1996: Modeling of land-surface evaporation by four schemes and comparison with FIFE observations. J. Geophys. Res, 101, 7251–7268

Choi M, Kustas WP, Anderson MC, Allen RG, Li F, & Kjaersgaard JH (2009). An intercomparison of three remote sensing-based surface energy balance algorithms over a corn and soybean production region (Iowa, US) during SMACEX. Agricultural and Forest Meteorology, 149(12), 2082–2097.

Cleugh HA, Leuning R, Mu Q, & Running SW (2007). Regional evaporation estimates from flux tower and MODIS satellite data. Remote Sensing of Environment, 106(3), 285–304.

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, … & Vitart F (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q.J.R. Meteorol. Soc, 137, 553–597. doi: 10.1002/qj.828

Döll P, Hoffmann-Dobrev H, Portmann FT, Siebert S, Eicker A, Rodell M, … & Scanlon BR (2012). Impact of water withdrawals from groundwater and surface water on continental water storage variations. Journal of Geodynamics, 59, 143–156.

Falcone JA, Carlisle DM, Wolock DM, & Meador MR (2010). GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. Ecology, 91(2), 621–621.

Ferguson CR, Sheffield J, Wood EF, & Gao H (2010). Quantifying uncertainty in a remote sensing-based estimate of evapotranspiration over continental USA. International Journal of Remote Sensing, 31(14), 3821–3865.

French AN, Jacob F, Anderson MC, Kustas WP, Timmermans W, Gieske A, … & Prueger J (2005). Surface energy fluxes with the Advanced Spaceborne Thermal Emission and Reflection radiometer (ASTER) at the Iowa 2002 SMACEX site (USA). Remote Sensing of Environment, 99(1–2), 55–65.

Gao H, Tang Q, Ferguson CR, Wood EF, & Lettenmaier DP (2010). Estimating the water budget of major US river basins via remote sensing. International Journal of Remote Sensing, 31(14), 3955–3978.

Glenn EP, Doody TM, Guerschman JP, Huete AR, King EA, McVicar TR, … & Zhang Y (2011). Actual evapotranspiration estimation by ground and remote sensing methods: the Australian experience. Hydrological Processes, 25(26), 4103–4116.

Haddeland I, Clark DB, Franssen W, Ludwig F, Voß F, Arnell NW, … & Gomes S (2011). Multimodel estimate of the global terrestrial water balance: setup and first results. Journal of Hydrometeorology, 12(5), 869–884.

Hain CR, Crow WT, Anderson MC, & Yilmaz MT (2015). Diagnosing Neglected Soil Moisture Source–Sink Processes via a Thermal Infrared–Based Two-Source Energy Balance Model. Journal of Hydrometeorology, 16(3), 1070–1086.

Han E, Crow WT, Hain CR, & Anderson MC (2015). On the use of a water balance to evaluate interannual terrestrial ET variability. Journal of Hydrometeorology, 16(3), 1102–1108.

Hijmans RJ, & van Etten J (2014). raster: Geographic data analysis and modeling. R package version, 2, 15.

Jiménez C, Prigent C, Mueller B, Seneviratne SI, McCabe MF, Wood EF, … & Fisher JB (2011). Global intercomparison of 12 land surface heat flux estimates. Journal of Geophysical Research: Atmospheres, 116(D2).

Jung M, Reichstein M, & Bondeau A (2009). Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. Biogeosciences, 6(10), 2001–2013.

Kalma JD, McVicar TR, & McCabe MF (2008). Estimating land surface evaporation: A review of methods using remotely sensed surface temperature data. Surveys in Geophysics, 29(4–5), 421–469.

Kustas WP, Alfieri JG, Anderson MC, Colaizzi PD, Prueger JH, Evett SR, … & Copeland KS (2012). Evaluating the two-source energy balance model using local thermal and surface flux observations in a strongly advective irrigated agricultural area. Advances in Water Resources, 50, 120–133.

Kustas WP, Norman JM, Hain CR, Mecikalski JR, Schultz L, González-Dugo MP, … & Gao F (2011). Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. Hydrology and Earth System Sciences, 15(1), 223.

Landerer FW, & Swenson SC (2012). Accuracy of scaled GRACE terrestrial water storage estimates. Water resources research, 48(4).

Li ZL, Tang R, Wan Z, Bi Y, Zhou C, Tang B, … & Zhang X (2009). A review of current methodologies for regional evapotranspiration estimation from remotely sensed data. Sensors, 9(5), 3801–3853. [PubMed: 22412339]

Liu W, Wang L, Zhou J, Li Y, Sun F, Fu G, … & Sang YF (2016). A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. Journal of Hydrology, 538, 82–95.

Lo MH, Swenson S, Famiglietti JS, Tang Q, Skaggs TH, Lin YH, & Wu RJ (2015). Using satellite-based estimates of evapotranspiration and groundwater changes to determine anthropogenic water fluxes in land surface models. Geoscientific Model Development, 8(10), 3021.

Mu Q, Heinsch FA, Zhao M, & Running SW (2007). Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. Remote sensing of Environment, 111(4), 519–536.

Mu Q, Zhao M, & Running SW (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sensing of Environment, 115(8), 1781–1800.

Newman AJ, Clark MP, Sampson K, Wood A, Hay LE, Bock A, … & Hopson T (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences, 19(1), 209.

Pan M, Sheffield J, Wood EF, Mitchell KE, Houser PR, Schaake JC, … & Luo L (2003). Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent. Journal of Geophysical Research: Atmospheres, 108(D22).

Peters-Lidard CD, Kumar SV, Mocko DM, & Tian Y (2011). Estimating evapotranspiration with land data assimilation systems. Hydrological Processes, 25(26), 3979–3992.

Pervez MS, & Brown JF (2010). Mapping irrigated lands at 250-m scale by merging MODIS data and national agricultural statistics. Remote Sensing, 2(10), 2388–2412.

Pinker RT, Tarpley JD, Laszlo I, Mitchell KE, Houser PR, Wood EF, … & Sheffield J (2003). Surface radiation budgets in support of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data

Assimilation System (NLDAS) project. Journal of Geophysical Research: Atmospheres, 108(D22).

Prat OP, & Nelson BR (2015). Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). Hydrology and Earth System Sciences, 19(4), 2037–2056.

Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E, … & Bloom S (2011). MERRA: NASA's modern-era retrospective analysis for research and applications. Journal of climate, 24(14), 3624–3648.

Rodell M, & Famiglietti JS (2001). An analysis of terrestrial water storage variations in Illinois with implications for the Gravity Recovery and Climate Experiment (GRACE). Water Resources Research, 37(5), 1327–1339.

Rodell M, Famiglietti JS, Chen J, Seneviratne SI, Viterbo P, Holl S, & Wilson CR (2004). Basin scale estimates of evapotranspiration using GRACE and other observations. Geophysical Research Letters, 31(20).

Swenson SC (2012). GRACE monthly land water mass grids NETCDF RELEASE 5.0. Physical Oceanography Distributed Active Archive Center (PO. DAAC), California, 10.5067/TELND-NC005.

Sahoo AK, Pan M, Troy TJ, Vinukollu RK, Sheffield J, & Wood EF (2011). Reconciling the global terrestrial water budget using satellite remote sensing. Remote Sensing of Environment, 115(8), 1850–1865.

Save H, Bettadpur S, and Tapley BD (2016), High resolution CSR GRACE RL05 mascons, J. Geophys. Res. Solid Earth, 121, doi:10.1002/2016JB013007.

Sheffield J, Ferguson CR, Troy TJ, Wood EF, & McCabe MF (2009). Closing the terrestrial water budget from satellite remote sensing. Geophysical Research Letters, 36(7).

Shiklomanov IA (2000). Appraisal and assessment of world water resources. Water international, 25(1), 11–32.

Trenberth KE, Fasullo JT, & Kiehl J (2009). Earth's global energy budget. Bulletin of the American Meteorological Society, 90(3), 311–323.

Twine TE, Kucharik CJ, & Foley JA (2004). Effects of land cover change on the energy and water balance of the Mississippi River basin. Journal of Hydrometeorology, 5(4), 640–655.

Velpuri NM, Senay GB, Singh RK, Bohms S, & Verdin JP (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 RK, Meynadier R, Sheffield J, & Wood EF (2011). Multi-model, multi-sensor estimates of global evapotranspiration: climatology, uncertainties and trends. Hydrological Processes, 25(26), 3993–4010.

Wang S, Pan M, Mu Q, Shi X, Mao J, Brümmer C, … & Black TA (2015). Comparing evapotranspiration from Eddy covariance measurements, water budgets, remote sensing, and land surface models over Canada a, b. Journal of Hydrometeorology, 16(4), 1540–1560.

Wiese DN (2015). GRACE monthly global water mass grids NETCDF RELEASE 5.0. Ver. 5.0. PO.DAAC, CA, USA Dataset accessed [2017-03-18] at 10.5067/TEMSC-OCL05.

Wheeler D, & Tiefelsdorf M (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical Systems, 7(2), 161–187.

Xia Y, Hobbins MT, Mu Q, & Ek MB (2015). Evaluation of NLDAS-2 evapotranspiration against tower flux site observations. Hydrological processes, 29(7), 1757–1771.

Xia Y, Mitchell K, Ek M, Sheffield J, Cosgrove B, Wood E, … & Livneh B (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. Journal of Geophysical Research: Atmospheres, 117(D3).

Yang D, Kane D, Zhang Z, Legates D, & Goodison B (2005). Bias corrections of long-term (1973–2004) daily precipitation data over the northern regions. Geophysical Research Letters, 32(19).

Zeng Z, Wang T, Zhou F, Ciais P, Mao J, Shi X, & Piao S (2014). A worldwide analysis of spatiotemporal changes in water balance-based evapotranspiration from 1982 to 2009. Journal of Geophysical Research: Atmospheres, 119(3), 1186–1202.

Zhang Y, Leuning R, Chiew FH, Wang E, Zhang L, Liu C, … & Jung M (2012). Decadal trends in evaporation from global energy and water balances. Journal of Hydrometeorology, 13(1), 379–391.

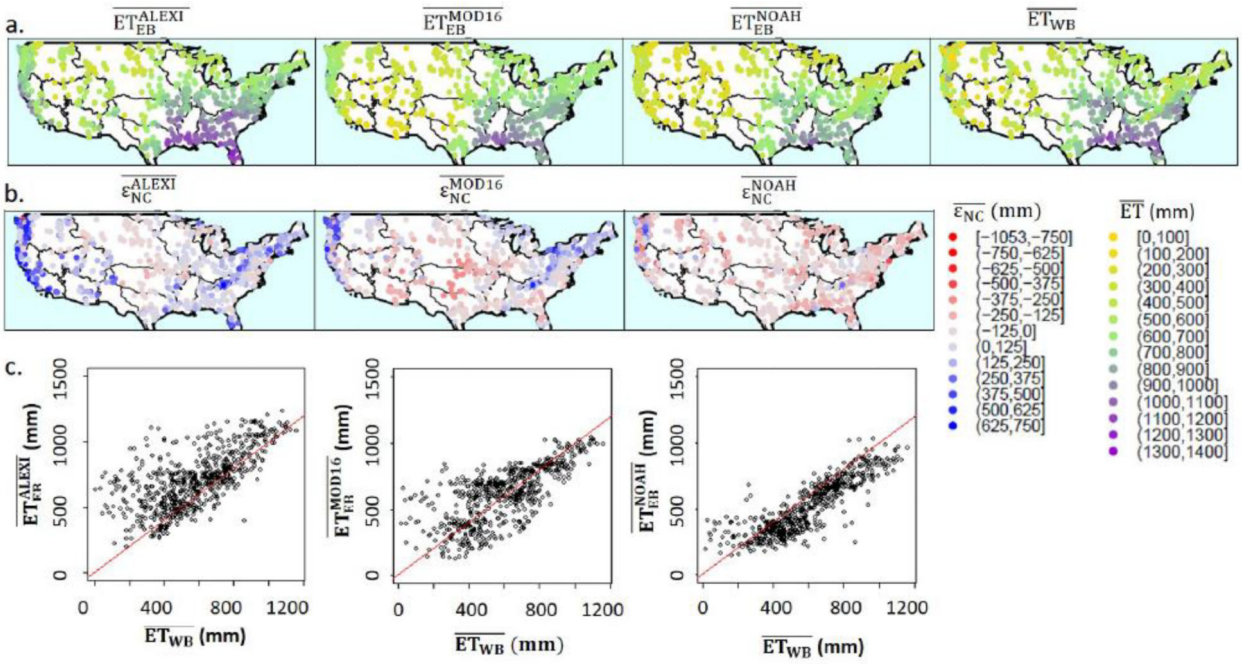

Figure 1:

a: Spatial distribution of long-term mean annual ET from terrestrial water balance and energy balance models. Color scale is in units of mm/year. Each point represents a streamflow gaging station. b: $\overline{\varepsilon_{NC}^{NOAH}}$, $\overline{\varepsilon_{NC}^{MOD16}}$, and $\overline{\varepsilon_{NC}^{ALEXI}}$ in mm/year, where positive $\overline{\varepsilon_{NC}}$ (blue) indicates that $\overline{ET_{EB}}$ overpredicts relative to $\overline{ET_{WB}}$. c: Scatterplot of $\overline{ET_{EB}}$ versus $\overline{ET_{WB}}$ with 1-to-1 line shown. Regular and square brackets represent inclusive and exclusive bounds of the interval, respectively.

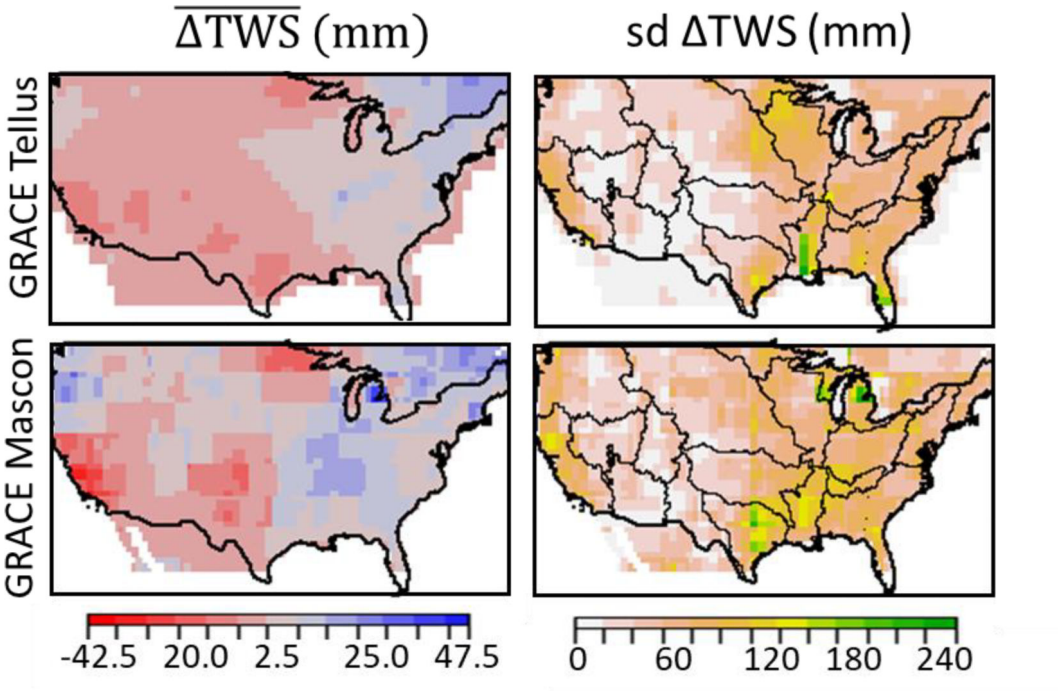

Figure 2:

Left column: 2003–2015 mean Δ*T*WS (in mm) calculated using GRACE Tellus (top) and GRACE Mascon (bottom) data products. Non-zero values indicate long term change in Sep-Oct annual water storage, with positive (blue) values indicating net gain in ΔTWS and negative (red) values indicating net loss in $\overline{\Delta TWS}$. Right column: standard deviation of 2003–2015 ΔTWS (in mm) calculated using GRACE Tellus (top) and GRACE Mason (bottom) data products.

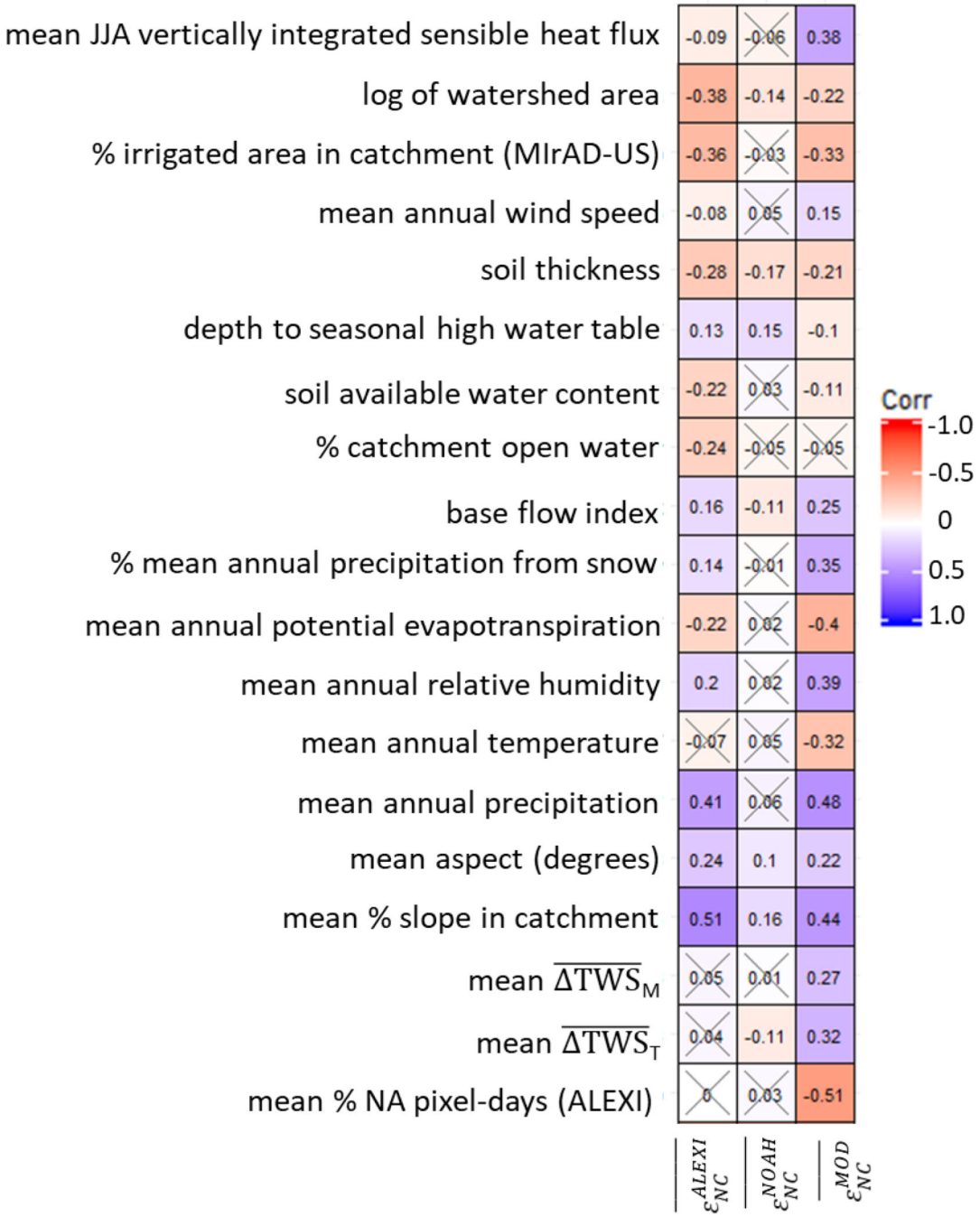

Figure 3:

Spearman's correlation coefficient between $\overline{\varepsilon_{NC}^{ALEXI}}$ (left column), $\overline{\varepsilon_{NC}^{NOAH}}$(center column) and $\overline{\varepsilon_{NC}^{MOD16}}$ (right column) and watershed characteristics. Color scale represents strength of negative (red) or positive (blue) correlation between non-closure error and watershed characteristic. Correlation coefficients which were not significant at α =0.05 are indicated with an X.

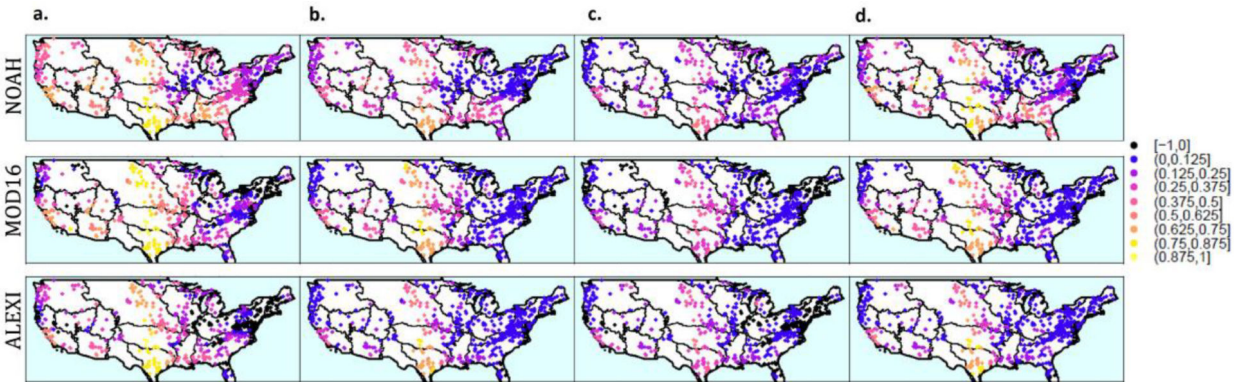

Figure 4:

a: Geographically weighted β estimates and b: local R^2 for centered and scaled data (Eqn. 3.1), where each dot represents a basin. c: Geographically weighted β estimates and d: local R^2 for unscaled data (Eqn. 3.2). Outlines represent USGS HUC02 Hydrologic regions across CONUS.

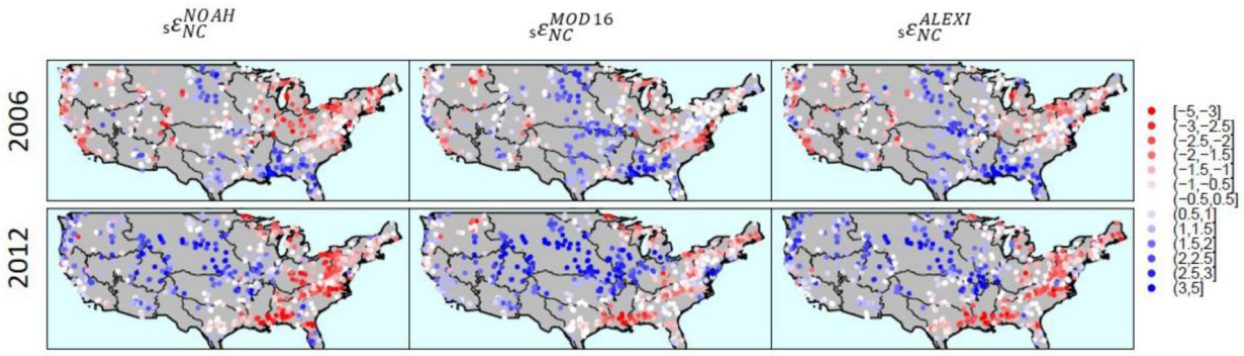

Figure 5:

2006 and 2012 spatial patterns in $_s\varepsilon_{NC}^{NOAH}$ (left), $_s\varepsilon_{NC}^{MOD16}$(middle), and $_s\varepsilon_{NC}^{ALEXI}$ (right).

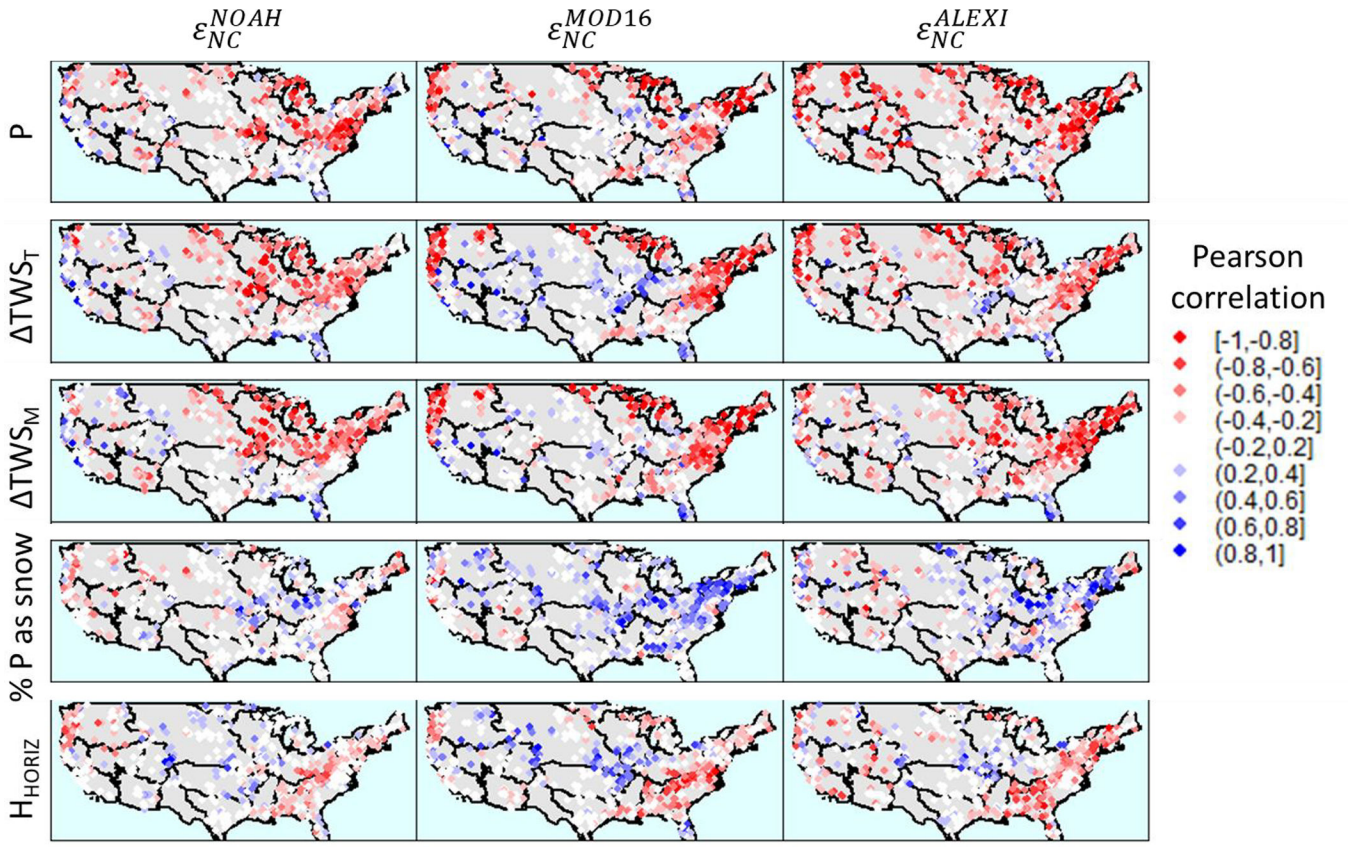

Figure 6:

Basin specific correlations between ${}_s\varepsilon_{NC}^{NOAH}$ (left), ${}_s\varepsilon_{NC}^{MOD16}$(center), and ${}_s\varepsilon_{NC}^{ALEXI}$ (right), and annual P (top row), annual ΔTWS_T (second row), annual ΔTWS_M (third row), % P as snow (fourth row) and JJA vertically integrated sensible heat flux (H_{HORIZ}, bottom row).