Implications of Model Selection: A Comparison of Publicly Available, CONUS-Extent Hydrologic Component Estimates

Samuel Saxe¹,², William Farmer¹, Jessica Driscoll¹, Terri S. Hogue²,³

²Hydrologic Science and Engineering, Colorado School of Mines, Golden, CO 80401, U.S.
³Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO 80401, U.S.

Correspondence to: Samuel Saxe (ssaxe@usgs.gov)

Abstract. Spatiotemporally continuous estimates of the hydrologic cycle are often generated through hydrologic modeling, reanalysis, or remote sensing methods, and commonly applied as a supplement to, or a substitute for, in-situ measurements when observational data are sparse or unavailable. This study compares estimates of precipitation (P), actual evapotranspiration (ET), runoff (R), snow water equivalent (SWE), and soil moisture (SM) from 87 unique datasets generated by 47 hydrologic models, reanalysis datasets, and remote sensing products across the conterminous United States (CONUS). Uncertainty between hydrologic component estimates was shown to be high in the western CONUS, with median uncertainty (measured as the coefficient of variation) ranging from 11-21% for P, 14-26% for ET, 28-82% for R, 76-84% for SWE, and 36-96% for SM. Uncertainty between estimates was lower in the eastern CONUS, with medians ranging from 5-14% for P, 13-22% for ET, 28-82% for R, 53-63% for SWE, and 42-83% for SM. Inter-annual trends in estimates from 1982-2010 show common disagreement in R, SWE, and SM. Correlating fluxes and stores against remote sensing-derived products shows poor overall correlation in the western CONUS for ET and SM estimates. Study results show that disagreement between estimates can be substantial, sometimes exceeding the magnitude of the measurements themselves. The authors conclude that multi-model ensembles are not only useful, but are in fact a necessity, to accurately represent uncertainty in research results. Spatial biases of model disagreement values in the western United States show that targeted research efforts in arid and semi-arid water-limited regions are warranted, with the greatest emphasis on storage and runoff components, to better describe complexities of the terrestrial hydrologic system and reconcile model disagreement.

1 Introduction

A long-term goal of the atmospheric and hydrologic scientific communities has been to produce accurate estimates of the hydrologic cycle across continental and global domains (Archfield et al., 2015; Beven, 2006; Freeze and Harlan, 1969). Various methodologies have been applied to meet this goal, typically in the form of physically based, reanalysis, or remote sensing-based models. Many of the resulting estimates are made publicly available by an assortment of scientific entities at both continental- and global-extents, across a wide range of spatiotemporal resolutions. These datasets accelerate progress in
the atmospheric and hydrologic sciences by filling knowledge gaps in data-sparse regions, reducing the need for time-consuming and computationally expensive modeling, and by providing numerous estimates to apply within ensemble analyses.

Publicly available modeled estimates have been applied to work on water budget analyses (Gao et al., 2010; Pan et al., 2012; Rodell et al., 2015; Smith and Kummerow, 2013; Velpuri et al., 2019; Zhang et al., 2018), effects of climate change (LaFontaine et al., 2015; G. J. McCabe et al., 2017), and water availability and use (Landerer and Swenson, 2012; Thomas and Famiglietti, 2019; Voss et al., 2013; Zaussinger et al., 2019). Rapid increases in computational power and data accessibility following the advent of early global- and continental-extent hydrologic models in the 1980s and 1990s (Koster and Suarez, 1992; Manabe, 1969; Sellers et al., 1986; Yang and Dickinson, 1996), and increasingly higher-resolution passive and active satellite measurements of both the surface and subsurface of the earth (Alsdorf et al., 2007; M. F. McCabe et al., 2017), have led to an explosion in the production of multi-decadal, continental- to global-extent models estimating all major components of the hydrologic cycle (Peters-Lidard et al., 2018). Accordingly, this has given rise to a multitude of model comparison and evaluation projects throughout the scientific literature.

Earlier model intercomparison projects, such as the Coupled Model Intercomparison Project (CMIP) (Covey et al., 2003), the Integrated Hydrologic Model Intercomparison Project (IH-MIP) (Kollet et al., 2017; Maxwell et al., 2014), the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013), and the Water Model Intercomparison Project (WaterMIP) (Haddeland et al., 2011), provide robust analyses aimed at attributing differences in model process representation to differences in model design, parameter selection, and meteorological forcings.

Detailed comparisons between smaller selections of models are often included in studies presenting novel products (Daly et al., 2008; G.B. Senay et al., 2011; N. M. Velpuri et al., 2013; Youlong Xia, Mitchell, Ek, Cosgrove, et al., 2012a; Youlong Xia, Mitchell, Ek, Sheffield, et al., 2012b). Further, a range of studies have focused solely on validation of third-party models, either examining multiple water budget components simultaneously (Gao et al., 2010; Sheffield et al., 2009) or focusing on specific components such as precipitation (P) (Derin & Yilmaz, 2014; Donat et al., 2014; Guirguis & Avisar, 2008; Prat & Nelson, 2015; Sun et al., 2018), evapotranspiration (ET) (Carter et al., 2018; M. F. McCabe et al., 2016), snow water equivalent (SWE) (Broxton et al., 2016; Chen et al., 2014; N. Dawson et al., 2018; Nicholas Dawson et al., 2016; Essery et al., 2009; Mudryk et al., 2015; Rutter et al., 2009; Vuyovich et al., 2014), or soil moisture (SM) (Brocca et al., 2011; Koster et al., 2009; Youlong Xia et al., 2014).

Most of these comparison studies have revolved around error metrics derived through validation against in-situ and ex-situ measurements. However, validation often yields unsatisfactory representations of model skill. Observational datasets are often spatiotemporally discontinuous and model estimates in grid structure may not sufficiently represent the sub-grid heterogeneity that is present in point data, especially in topographically and ecologically complex regions. Further, observational
measurements can have significant associated uncertainty (Baldassarre and Montanari, 2009). Validation literature is heavily focused on comparison of error statistics and spatial summaries of model skill, rarely discussing the effect of model disagreement on research results.

This research seeks to supplement past comparison and validation literature by evaluating how uncertainty between publicly available water budget component estimates used within a study can control outcomes and conclusions across a range of ecological regimes. We differentiate this work from previous MIPs, MFJs, and CSAs by placing results in the context of multiple water budget analyses. We seek to quantify how estimates of component magnitudes and long-term trends differ in the CONUS, as well as within ecologically distinct regions. The primary goal of this work is to quantify model disagreement in terms of magnitude, inter-annual trend, and correlation against remote sensing products. The effects of model disagreement are shown in regional water budget analyses. We undertake a robust comparison of P, ET, R, SWE, and SM estimates generated through hydrologic models, reanalysis datasets, and remote sensing products. We iteratively calculate water budget imbalances in eight regions, by applying a range of flux estimates to quantify how model selection may impact residuals.

2 Methods and Data

2.1 Data Categories

Hydrologic estimates are divided into water budget flux components of P, ET, and R, and storage components of SWE and SM), representing the primary fluxes and stores of a surface water budget. Datasets were selected by prioritizing public availability, ease of access, and relative use within the research community. These datasets are subdivided into loosely defined categories of hydrologic models, reanalysis datasets, and RS-derived products. References and spatiotemporal information for each dataset are provided in Table 1 and more detailed long-form descriptions are provided in Appendix 1.

2.1.1 Hydrologic Models

The hydrologic model category (Table 1, Appendix 1) includes any estimates generated using equations or concepts attempting to represent real-world hydrology. Model output differences are strongly controlled by forcing datasets (Elsner et al., 2014; Mizukami et al., 2014), calibration methods (Mendoza et al., 2015), applied equations (Clark et al., 2015a), model structure (Clark et al., 2015a, 2015b), and geophysical parameter availability (Beven, 2002; Bierkens, 2015; Mizukami et al., 2017). The most common hydrologic models this study evaluates are conceptual and physically based. Conceptual models derive terrestrial hydrology estimates through empirical relationships between fluxes and stores, typically through a water balance model (WBM) in the style of Thornthwaite (1948) (NHM-MWBM and TerraClimate). Physically based models utilize meteorological forcing data and solve equations describing physical conservation laws of mass, energy, and momentum. These can be further sub-divided by targeted hydrologic variables: land surface models (LSMs) target land-atmosphere interactions, especially ET, while catchment models (CMs) target streamflow (Archfield et al., 2015). Additionally, LSMs are often one-dimensional models that operate at discrete spatial intervals, typically grid cells, lacking horizontal transfer of surface or
subsurface water between regions. CMs, on the other hand, utilize two-dimensional model structures by routing overland and subsurface flows between spatial domains to better realize surface and groundwater estimates using prescribed stream networks. The CLSM, described in the literature as a land surface model (Koster et al., 2000), is discussed in conjunction with the NHM-PRMS because of the sub-grid catchment network used to model horizontal runoff and streamflow fluxes. LSMs are the most common hydrologic model here (e.g. CLM, H-TESSEL, Mosaic, Noah, SiB, and VIC) because they are often coupled with global-extent atmospheric research and thus more commonly operated at the spatial extent of this study, while CMs are more typically operated at the catchment or basin scale. Additional products included in the hydrologic model category are those using simplified WBM's falling outside the conceptual model paradigm (CPC, CSIRO-PML, GLEAM, and VegET) and those using component-specific physically based models (SNODAS).

2.1.2 Reanalysis

Reanalysis datasets (Table 1, Appendix 1) assimilate multi-source in-situ and ex-situ observational data into spatiotemporally continuous four-dimensional estimates of continental- or global-scale atmospheric and meteorological fluxes using numerical algorithms. This category includes reanalysis datasets derived solely from in-situ and ex-situ data (e.g. CMAP, DayMET, Maurer et al. 2002, UoD-v5), as well as those assimilating or blending multiple reanalysis products with or without observational measurements (e.g. CanSISE, gridMET, Livneh et al. 2013, NLDAS2). In past studies, reanalysis models were often grouped and defined separately from gridded datasets derived through statistical interpolation of in-situ observations (e.g. WaterWatch, Reitz et al. 2017, GPCC). We combine the two categories, defining the single reanalysis category as containing products not generated solely from remote sensing observations or through terrestrial hydrologic models. The reasons for this are two-fold: (1) to simplify reporting of results and discussion across water budget components, and (2) although interpolation-based products are often used as reference datasets in model validation studies, especially for precipitation, accuracy decreases in regions with sparse observations and complex topography (Hofstra et al., 2008) and so should be considered modeled products.

2.1.3 Remote Sensing-Derived

The RS-derived datasets (Table 1, Appendix 1) are components of the hydrologic system derived from passive and/or active ex-situ observations. We note the use of the term “RS-derived” rather than “RS” because none of the datasets used here can be truly described as direct observations, always relying on a range of modeling techniques to create hydrologic estimates. For example, the AMSR-E/Aqua SWE product uses a sequence of steps including a snow detection routine, followed by a physical retrieval algorithm, which is then validated against observational SWE data (Chang and Rango, 2000), and the SSEBop product uses a radiation-driven energy balance model to estimate ET (Senay et al., 2013). Remote sensing datasets include estimates modeled from single-sensor ex-situ observations (MOD16-A2, SSEBop), though most utilize an assembly of information from both passive and active sensors at various spatiotemporal resolutions (e.g. ESA-CCI, GPCP-v3).
2.2 Data Processing

Datasets included in this research cover a broad spectrum of spatial resolutions, extents, coordinate reference systems, and time scales, and therefore required a uniform spatiotemporal system to better facilitate comparison. To that end, gridded (e.g. NLDAS2-Mosaic) and polygonal (e.g. NHM-PRMS) datasets were aggregated by area-weighted mean to 10 Environmental Protection Agency Ecoregions, Level I (Omernik and Griffith, 2014) that encompass the Watershed Boundary Dataset Hydrologic Units (U.S. Geological Survey, 2016) (Fig. 1, Table 2) over the CONUS. Datasets were processed in their native coordinate system to avoid interpolation of raw values and the reference system of the ecoregions spatial dataset was transformed to match that of each target dataset. Flux and storage terms were aggregated to monthly time steps. Flux values (P, ET, and R) were summed from hourly or daily when necessary and storage values (SWE, SM) were averaged.

Units of P, ET, R, and SWE are uniformly presented as equivalent water depth in millimeters (mm). SM is provided by datasets as either equivalent water depth (mm) or volumetric soil moisture content (m3/m3), and denoted as SM(e) or SM(v), respectively. The SM(e) and SM(v) categories are merged during results and discussion when magnitude is irrelevant, such as with long-term trend and correlation against RS products. All data generated through these methods are available in an associated data release hosted on the USGS ScienceBase (Saxe et al., 2020).

2.3 Statistics

2.3.1 Annual Uncertainty

Uncertainties in hydrologic component values are measured as the variability between modeled estimates using coefficient of variation (CV) calculated as:

\[ CV = \frac{\sigma}{\bar{x}} \times 100\% \]

where \( \sigma \) is the standard deviation across all products of a given component in a given water year and \( \bar{x} \) is the associated mean. The \( \sigma \) and CV statistics were calculated for each water year (WY; a water year is the 12-month period October 1 through September 30 designated by the calendar year in which it ends) between available estimates for all components. Annual values of flux terms (P, ET, and R) were derived by summing monthly values to WYs, and storage terms were derived by averaging monthly values by WY. Incomplete WYs (n months < 12) were discarded. Because of incompatibility between dataset temporal ranges, \( \sigma \) and CV values were divided into two 15-WY periods: the early period consisting of WYs 1985-1999, and the late period consisting of WYs 2000-2014.

Uncertainty metrics are impacted by the number of models available for each calculation. That is, increasing the number of models typically increases the inter-model variability. Depending on water year, there can be 14-16 P models, 11-17 ET models, 12-15 R models, 14-20 SWE models, 7-9 SM(e) models, and 5-7 SM(v) models available. To reduce the effects of
model count disparity on variability measures, $CV$ and $\sigma$ are calculated through a bootstrap approach that calculates the $CV$ and $\sigma$ for all possible combinations of 5 models (from the minimum SM(v) availability) by water year, and returns final $CV$ and $\sigma$ as the mean of the bootstrapped combinations.

2.3.2 SWE Timing

Variability in seasonal timing of SWE estimates was compared for trends of both accumulation and ablation in terms of relative timing, defined here as the difference between the antecedent month of snow accumulation or ablation for each dataset from the mean antecedent month of the remaining datasets, calculated for each Julian year from 1985-2014. The antecedent accumulation (ablation) month is defined as the first month of the June-December (January-May) period, where change in SWE is greater than (less than) 1 mm per month (-1 mm per month). Some datasets contained anomalies in January and February, showing a large negative change in SWE followed by positive rates of SWE accumulation for 1-2 more months, which introduced a negative bias into relative timing values. To eliminate these biases, ablation timing was only considered for months subsequent to the final date of positive SWE rate of change. Resulting annual relative timing values are presented for each SWE product grouped by (a) early and late periods, (b) accumulation and ablation, and (c) ecoregions Northern Forests and Northwestern Forested Mountains. To simplify comparison of SWE timing between trends and ecoregions, annual directions of relative timing were summarized by percent direction, defined here as the percentage of years with a positive relative timing value:

$$p^+ = \frac{|R^+|}{|R|} \times 100\%,$$

where $p^+$ is percent direction, $|R^+|$ is the cardinality (number of elements in a set) of positive, real relative timing values, and $|R|$ is the cardinality of all real values. Within the figure presenting summary values (Appendix 2.4), percent direction values less than 50% (i.e. those dominantly negative) are converted to a scale of 50-100% with:

$$pD = \begin{cases} p^+ & p^+ > 50 \\ 100\% - p^+ & p^+ < 50 \end{cases},$$

where $pD$ is the directionalized percent direction colored according to the dominant direction value. Positive $pD$ uses a green gradient and negative $pD$ uses a purple gradient.

2.3.3 Inter-annual Trend

Both the Mann-Kendall trend test ($\tau$) (Kendall, 1938) and Sen’s slope estimator (Sen, 1968) are used to identify and measure monotonic trends in annual values over WYs 1982-2010. Trend significance is evaluated using p-values ($p$) in a binary significance test, assuming an alpha ($\alpha$) of 0.05 and a null hypothesis of no monotonic trend. Under the condition of $p < \alpha$, the null hypothesis is rejected and the alternative hypothesis, that of the presence of a monotonic trend, is assumed; if $p > \alpha$, the
null hypothesis is not rejected. Disagreement in the presence of significant trend and trend direction is quantified using the unalikeability coefficient \((u)\) which measures how often categorical variables differ on a \(0 \leq u \leq 1\) scale, with 0 and 1 being complete agreement and disagreement, respectively (Kader and Perry, 2007). This study compares trends between 11 to 16 different products \((11 \leq n \leq 16)\) depending on water budget component and uses categorical values \((c)\) of (a) significant negative trend, (b) significant positive trend, and (c) no significant trend, derived from the direction (sign) and significance \((\alpha = 0.05)\) of \(\tau\). When \(n > c\), maximum possible \(u\) decreases from 1 exponentially to an asymptote of 2/3 as the ratio of \(n:c\) increases. With a \(n:c\) ranging from 11:3 to 16:3, maximum possible \(u\) in this trend analysis is approximately 0.71.

Spearman’s rho (Spearman, 1904) is applied to quantify correlation between RS components \(P\), ET, SWE, and SM and the analogous components of hydrologic models and remote sensing datasets. Spearman’s rho \((\rho)\), a non-parametric rank correlation metric, is used to estimate correlation, producing values on a \(-1\) to 1 scale where 1 is perfect positive correlation and -1 is perfect negative correlation:

\[
\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},
\]

where \(d\) is the difference between ranks and \(n\) is sample size. A binary significance test is used to test correlation statistics, assuming \(\alpha = 0.05\) and a null hypothesis that there is no relationship between datasets. If \(p < \alpha\), the null hypothesis is rejected and a significant correlation between the datasets is assumed; if \(p > \alpha\), the null hypothesis is not rejected. Correlation is computed and assessed along the monthly timestep, requiring a minimum 48 months of temporal overlap between the modeled estimates and remote sensing dataset.

### 2.3.4 Water Budget Imbalances

Water budgets are calculated assuming a steady-state system and solved for imbalances by:

\[
P = AET + R + \varepsilon,
\]

\[
\varepsilon = P - AET - R,
\]

where \(\varepsilon\) is imbalances. Imbalances, \(\varepsilon\), cannot be accurately defined as residuals. In reality, \(\varepsilon\) is the sum of excluded fluxes, in addition to model uncertainty (residuals). Excluded fluxes include both natural (e.g. groundwater recharge, changes to long-term storage) and anthropogenic (e.g. groundwater extraction) hydrologic processes. Relative imbalances \((R\varepsilon)\) are calculated to better compare water budget results between regions of varying hydrologic flux by weighting water budget \(\varepsilon\) against the total input \(P\):

\[
R\varepsilon = (\varepsilon/P) \times 100\%.
\]
Water budget relative imbalances were calculated from summed P, ET, R, and ε over the 10-water year period of 2001-2010 for 15 P, 15 ET, and 13 R estimates (noted in Table 1) with temporally continuous monthly data for ten ecoregions. Each ecoregion yielded 2,925 water budgets by iterating through all possible combinations of models, totaling 29,250 water budgets in primary regions of the CONUS. In addition, a single water budget was calculated for each of the eight ecoregions using the ensemble means of the model estimates over the same period.

3 Results

3.1 CONUS Domain Average

3.1.1 Magnitude Variability

Uncertainties between annual estimates of states and fluxes of the hydrologic cycle are substantial when averaged over the CONUS for each studied component (Fig. 2). The actual magnitude of model estimate differences averaged by water budget component, measured as σ, is similar for P, ET, and R (Fig. 2a), and low (6.6 mm/yr) for SWE. However, comparing uncertainty relative to mean magnitude (Fig. 2b), measured as CV, shows that the vertical, atmospheric-controlled fluxes of P and ET demonstrate lower inter-model uncertainty than the horizontal flux of R or storage components SWE and SM. The CV statistic shows how, due to the low overall magnitude of SWE, small differences in estimates have a substantial effect on magnitude. SM, which is often a large overall storage term in the water budget, also shows high CV, indicating that uncertainty in estimates of this component may have the largest impact on hydrologic analyses.

Unfortunately, high-magnitude differences in SM (Fig. 3e & 3f) are likely strongly controlled by model-defined soil layer depth and thus limit the utility of disagreement statistics. For example, the TerraClimate and NLDAS2-Noah models apply spatially invariant rootzone soil depth across the model domains (Abatzoglou et al., 2018; Y. Xia et al., 2012a), whereas the NHM-PRMS assumes a variable depth according to vegetative and geophysical parameters (Regan et al., 2018). SM(e) estimates are directly controlled by soil depth, returning values of equivalent water depth. SM(v) differences are likely to be less influenced by soil depth due to inherent measurements of fractional volume rather than depth. However, soil-water profiles can change significantly with depth, so attributing model differences strictly by rootzone depth definition is more difficult.

Boxplots of modeled estimates by water budget component (Fig. 3) demonstrate the annual ranges of magnitude generated by various products. Precipitation models (Fig. 3a) exhibit low overall variability, with median magnitudes falling within a distinct 100 mm/yr (700-800 mm/yr) range. The exceptions to this are the CMAP reanalysis datasets, which have a median 647 mm/yr.

Variability among ET estimates (Fig. 3b) is higher, falling within a 200 mm/yr (435-650 mm/yr) range, though attributing differences to model type (e.g. LSMs vs. CMs) is difficult. Generally, CMs (MERRA-2/CLSM, MERRA-Land/CLSM, NHM-PRMS) and conceptual WBMs (NHM-MWBM, GLEAM, TerraClimate) produce greater annual rates of ET as compared to
LSMs (NLDAS2-VIC, GLDAS-CLM, NLDAS2-Noah). Exceptions to this are the LSMs NLDAS2-MOSAIC and H-TESSEL that are more similar in annual ET flux to CMs and WBMs. The RS datasets agree more at the CONUS-extent with lower-magnitude estimates.

Estimates of R show three distinct clusters of CONUS magnitudes (Fig. 3c) of 130-190, 220-242, and 310-345 mm/yr. The NLDAS2-Noah and NLDAS2-VIC LSMs, which produced some of the lowest ET rates, fall within the higher magnitude R clusters, as do the JRA-driven SiB models and NHM-PRMS model. The lowest magnitude estimates are the NLDAS2-Mosaic and ERA5-driven H-TESSEL LSMs, grouped with the CM MERRA-Land/CLSM and WBM TerraClimate. LSMs are more likely to estimate greater R than WBMs or CMs.

Boxplots of annually averaged SWE monthly values (Fig. 3d) show that WBMs, notably the NHM-MWMB, generate much higher SWE than most other datasets. Alternatively, there are few discernible patterns between the remaining products, with LSMs and CMs interspersed throughout the gamut of median values. To generalize, the LSMs Noah, GLDAS-CLM, and NLDAS2-Mosaic produce lower-magnitude estimates of monthly SWE. Three different Noah estimates, driven with different meteorological forcings and run both independently (NLDAS2-Noah) or as part of a larger model (e.g. NCEP-NARR/Eta-Noah), all estimate lower SWE relative to other datasets. The two VIC LSMs show contrasting median monthly SWE magnitude over the CONUS, with the Livneh-VIC median exceeding the NLDAS2-VIC median by more than 200%. The RS AMSR-E/Aqua product agrees more with lower-magnitude estimates at the CONUS extent.

Across all water budget components, most datasets demonstrate lower-magnitude values in the late period (WYs 2000-2014) compared to the early period (WYs 1985-1999). Within the P flux, almost all datasets agree on a decrease in magnitude of hydrologic fluxes from early to late periods. This is similarly reflected in the flux estimates of ET and R, and, to a less uniform extent, the storage components of SWE and SM. Several datasets, however, exhibit increased estimates from the early to late periods. For example, the NCEP-DOE reanalysis precipitation, and corresponding runoff derived by forcing the Noah LSM within the Eta atmospheric model, show increased median annual flux rates from the early to late periods.

3.1.2 Trends

Only 19 out of 87 component estimates exhibit statistically significant CONUS average domain trends, evaluated with Sen’s slope, from 1982 to 2010 (Fig. 4). Within most components, one or more datasets produced a significant slope that is contradicted by one or more datasets. For example, the R estimates from LSMs GLDAS-CLM, JRA-55/SiB, and ERA5/H-TESSEL each show a significant negative trend in annual values across the CONUS over the 1982-2010 period. Conversely, the NCEP-DOE/Eta-Noah LSM shows a significant positive trend over the same period and the remaining datasets, a mix of hydrologic and reanalysis models, show no significant trend. While the NCEP-DOE reanalysis precipitation trend matches that of the NCEP-DOE forced Eta-Noah, model estimates of SWE indicate a significant negative trend.
3.2 Ecoregions

3.2.1 Magnitude Variability

Uncertainty in modeled hydrologic component estimates in each ecoregion is presented in terms of CV (Fig. 5) and \( \sigma \) (Appendix 1). Ecoregions are organized from west to east within each figure. Results and discussion of uncertainty are focused primarily on CV due to the disparity in annual water flux between regions. By component, P and ET estimates demonstrate lower uncertainty compared to other hydrologic components (Fig. 5a & 5b). P estimates typically range in median uncertainty from 5-14% in the eastern CONUS and 11-21% in the western CONUS. P uncertainty is highest in mountainous regions with more variable topography (e.g. Northwestern Forested Mountains and Temperate Sierras) and lowest in more humid, topographically homogeneous regions (e.g. Eastern Temperate Forests).

The median uncertainties of ET estimates typically range from 13-22% in the eastern CONUS ecoregions and 14-26% in the western CONUS. Uncertainty is greatest along the Pacific coast and northerly ecoregions, as well as arid and semi-arid ecoregions that are heavily influenced by anthropogenic water use. Of the four largest ecoregions (Table 2), Northwestern Forested Mountains and North American Deserts demonstrate mean uncertainties of 21% and 19%, respectively, and the Great Plains and Eastern Temperate Forests demonstrate mean uncertainties of 14% each.

Estimates of R (Fig. 5c) generally exceed 29% variability throughout the study ecoregions, with median values reaching or exceeding 72% in arid, snow-sparse western ecoregions of North American Deserts, Southern Semiarid Highlands, and Temperate Sierras, as well as in Tropical Wet Forests influence by tropical storm systems. In the large Northwestern Forested Mountains and Great Plains ecoregions, median dataset variability is 28-55% and 47-81%, respectively.

SWE estimate uncertainty is highlighted for the snow-dominated Northwestern Forested Mountains and Northern Forests ecoregions (Fig. 5d) where annual peak SWE values exceed estimated snowpack in the remaining ecoregions by up to 700% (Appendix 2.2). Of primary interest is the Northwestern Forested Mountains ecoregion that contains the snowmelt dominated regimes of the Rocky, Sierra Nevada, and Cascade mountain ranges that are strong contributors to water supply for population centers such as Denver, Los Angeles, San Francisco, Portland, and Seattle. In this ecoregion, modeled estimates of monthly mean SWE storage vary by 76-84%, equating to median equivalent water depth standard deviations as high as 48 mm/month.

SM(e) estimate uncertainty (Fig. 5e) is greatest, 66-96%, in arid western ecoregions, though other regions fall within a 64-83% uncertainty range. SM(v) shows lower overall uncertainty (Fig. 5f) with less dramatic differences between regions, ranging from 36-57%, with the greatest uncertainty (44-57%) in central and eastern CONUS ecoregions. As mentioned previously, uncertainty in SM estimates, most importantly SM(e), is strongly influenced by model-defined rootzone soil thickness. Spatial differences in SM(e) uncertainty measured as CV, therefore, are pronounced in regions with lower water content because CV, a measure of relative variability, is more strongly affected by magnitude differences. Because the range
of values is constrained from 0-1 in m3/m3, uncertainty in SM(v) is less influenced by spatial variability in soil water magnitudes and is therefore a better measure for understanding regional soil water estimate disagreement. Following this logic, SM(v) uncertainty is greater in Tropical Wet Forests, Northern Forests, and Great Plains ecoregions with an average CV of 53%, compared to an average of 42% for the remaining ecoregions.

3.2.2 SWE Accumulation and Ablation

Timing of SWE estimates is presented in terms of relative timing (Appendix 2.3 & 2.4). Figures are divided by Northern Forests and Northwestern Forested Mountains ecoregions, trends of accumulation and ablation, and early and late periods. Negative (positive) values in purple (green), identify models where accumulation or ablation begins more than 1 month earlier (later) than the mean of other datasets. Distributions of relative timing by year are presented in Appendix 2.3 and summarized as percentage of years that are positive or negative in Appendix 2.4. In both ecoregions, uncertainty among datasets is greater in ablation timing than accumulation timing, and overall uncertainty is greater in the Northwestern Forested Mountains ecoregion than in the Northern Forests.

Relative timing between models can be similar between ecoregions (Appendix 2.3a, 2.4), especially regarding accumulation. For instance, the AMSR-E/Aqua RS and NHM-MWBM CM consistently show later accumulation start dates than other datasets in both ecoregions, while the ERA5-Land/H-TESSEL, JRA-55/SiB, and SNODAS models estimate earlier accumulation dates. Conversely, the JRA-25/SiB and TerraClimate models show different trends in accumulation timing between regions with earlier SWE accumulation in Northern Forests but later accumulation in Northwestern Forested Mountains.

Regarding ablation, outliers are much more common when comparing models (Appendix 2.3b, 2.4). For example, the DayMET model consistently estimates a much later start of spring ablation while the Eta-Noah models driven with NCEP-DOE and NCEP-NARR reanalyses typically estimate much earlier ablation times, often 1-2 months after the mean antecedent month (Appendix 2.3). Uncertainty in ablation timing is typically much higher in the Northwestern Forested Mountains ecoregion than in the Northern Forests. The DayMET, ERA5-Land/H-TESSEL, Livneh-VIC, NLDAS2-VIC, and SNODAS models commonly estimate a later start to ablation than the remaining models.

Models that predict earlier (later) accumulation timing and later (earlier) ablation timing represent longer (shorter) periods of growing or available snowpack (Appendix 2.4). The ERA5-Land/H-TESSEL, Livneh-VIC, and SNODAS models estimate longer snowpack periods than other models in both ecoregions. The AMSR-E/Aqua, JRA-55/SiB, NHM-PRMS, and TerraClimate models estimate longer snowpack periods in only one ecoregion. The NCEP-NARR/Eta-Noah, NLDAS2-Mosaic, and NLDAS2-Noah models estimate shorter snowpack periods in the Northern Forests ecoregion while the GLDAS-CLM, JRA-25/SiB models estimate shorter snowpack periods in the Northwestern Forested Mountains ecoregion.
Attributing relative timing of SWE to model type is difficult at the monthly scale. LSMs estimate both earlier and later accumulation and ablation antecedences, and the two WBM s in this study show opposing timing trends. Similarly, grouping by organization does not yield significant similarities. For example, the timing of NLDAS2-driven LSMs from the National Aeronautics and Space Administration (NASA) is dissimilar from that of the MERRA-2- and MERRA-Land-driven CLSM, also from NASA. Forcing data also does not yield useful information for identifying controlling variables. Only the NLDAS2-driven models from NASA tend to all show later (earlier) accumulation (ablation) timing. Even the Eta-Noah LSMs, driven by the NCEP-DOE reanalysis and the corresponding higher-resolution version for North America NCEP-NARR, show different relative timing values in both ecoregions.

### 3.2.3 Trend

Inter-annual trend analyses performed with the Mann-Kendall trend test ($\tau$) over water years 1982-2010 show varying degrees of agreement and disagreement between ecoregions. Figures showing explicit distributions of trends with model names are supplied in the appendix (Appendix 2.5-2.9). Model disagreement in trend direction is quantified using the unalikeability coefficient ($u$) (Fig. 6), with 0 and 1 being complete agreement and disagreement, respectively.

Precipitation estimates demonstrate the lowest $u$, showing no ($u = 0$) to low ($u = 0.13$) disagreement in eight ecoregions where datasets show no significant $\tau$ (Appendix 2.5). Disagreement is higher in North American Deserts and Temperate Sierras ecoregions ($u = 0.41-0.49$) where most datasets show a negative $\tau$, and all datasets show a negative $\tau$ in the Southern Semiarid Highlands. Coefficient $u$ is higher among ET datasets than P datasets, averaging 0.21 across ecoregions, notably in the North American Deserts and Marine West Coast Forest ecoregions ($u = 0.46$ and 0.53) where negative and insignificant $\tau$s are present (Appendix 2.6). All datasets identify negative ET $\tau$ in the Southern Semiarid Highlands and Temperate Sierras. Datasets show positive, negative, and insignificant $\tau$ in the Eastern Temperate Forests ($u = 0.31$) and Northern Forests ($u = 0.53$). Runoff datasets show the most consistent spatial distribution of $u > 0$ across the study ecoregions. Disagreement in western ecoregions is caused by conflicting negative and insignificant $R$ $\tau$ (Appendix 2.7), with the greatest $u$ in the Southern Semiarid Highlands ($u = 0.50$) and Temperate Sierras ($u = 0.49$). Eastern ecoregion $\tau$ is generally caused by conflicting positive, negative, and insignificant $\tau$s ($u = 0.26-44$). Ecoregions with good agreement ($u = 0-0.35$) is generally due to most datasets showing no significant trend.

SWE datasets, limited in this study to the Northwestern Forested Mountains and Northern Forests, show disagreement, $u = 0.42$ and 0.40, respectively, mostly caused by conflicting negative and insignificant $\tau$, though most datasets show no significant trend (Appendix 2.8). Trend agreement is better among SM datasets (mean $u = 0.25$), though higher $u$ is noted in the Northwestern Forested Mountains, North American Deserts, and Northern Forests, caused by conflicting negative and insignificant $\tau$s. Ecoregions with low $u$ typically show no significant $\tau$s in SM datasets except for the Southern Semiarid Highlands where most datasets show a significant negative $\tau$.
Generally, trend disagreement is highest in North American Deserts and Northern Forests ecoregions and lowest in Mediterranean California, Eastern Temperate Forests, Marine West Coast Forest, and Tropical Wet Forests ecoregions. Trend disagreement is almost always caused by conflicting negative and insignificant trends, indicating that disagreement is due to occurrence and not direction of trend. That is to say, higher u values are caused by disagreement over whether there is or is not a significant trend present in the data, as opposed to higher u values being caused by disagreement between significant negative and significant positive trends.

### 3.2.4 Correlation with Remote Sensing

Spearman’s rho (ρ) correlation was calculated for all hydrologic models and reanalysis datasets against remote sensing products with 48 or more months of intersecting temporal extents (Fig. 7, Table 3). Precipitation datasets (Fig. 7a & b), correlated against the GPCP-v3 and TMPA-3B43 products, show high correlation compared against 13 datasets (ρ = 0.93-0.99) with no statistically insignificant values. Variability of P estimate correlations are low (± 0.01-0.09). ET datasets correlated against the MOD16-A2 and SSEBop products (Fig. 7c & d) show poorer correlation in western ecoregions, with mean ρ ranging from 0.59-0.91. Extreme cases are found when correlating against MOD16-A2 in North American Deserts (ρ = -0.28) and SSEBop in Mediterranean California (ρ = 0.25) where four datasets show insignificant ρ. The exception in the western CONUS is in Northwestern Forested Mountains where models correlate against both RS datasets well (MOD16-A2 ρ = 0.91 and SSEBop ρ = 0.89). Correlation against both products is high in eastern ecoregions, ranging from ρ = 0.88-0.96 with standard deviation decreasing as ρ increases.

SWE datasets correlated against the AMSR-E/Aqua product are highlighted for the Northwestern Forested Mountains and Northern Forests ecoregions (Fig. 7e). Modeled estimates correlate better in the Northwestern Forested Mountains (ρ = 0.91) than Northern Forests (ρ = 0.80). SM datasets show worse and more variable ρ against remote sensing products ESA-CCI and SMOS-L4 (Fig. 7f & g) than other components. At the CONUS scale, estimates correlate very poorly against the ESA-CCI product (ρ = 0.17) and only moderately well against SMOS-L4 (ρ = 0.68). Generally, correlation is best in the Mediterranean California and Eastern Temperate Forests ecoregions (ρ = 0.74-0.92). Correlation is worst in Northwestern Forested Mountains, North American Deserts, and Northern Forests (ESA-CCI ρ = 0.44-0.61, SMOS-L4 ρ = 0.12-0.28). Explicit distributions of correlation for each dataset by ecoregion are provided in Appendix 2.10.

### 3.2.5 Case Study: Impact of Model Selection on Water Budget Imbalances

A case study calculating 2,925 10-year water budgets (WY 2001-2010) using 15 P, 15 ET, and 13 R estimates for each of ten ecoregions was performed to demonstrate quantitatively how model selection can affect research results. Each water budget was solved for a relative imbalance (\(R_e\)) with Eq. (7), estimating model error as a fraction of total water flux in P.
Histograms of results overlain with boxplots (Fig. 8) demonstrate the range and distribution of potential water budget $R_\text{es}$. In most ecoregions, $R_\text{es}$ exhibit an approximately normal distribution. Four ecoregions dominating the area of the CONUS (Fig. 8a-d), constituting 90% of the CONUS area (Table 2), are the focus of these results: the western ecoregions Northwestern Forested Mountains and North American Deserts, and the eastern ecoregions Great Plains and Eastern Temperate Forests (11%, 19%, 29%, and 31% of CONUS area, respectively).

Of these larger domains, water budget $R_\text{es}$’s in the eastern ecoregions show much lower variability (Table 4), with Great Plains and Eastern Temperate Forests yielding $\sigma$ of 18.2 percentage points and 14.3 percentage points, respectively, and medians of 3.9% and -0.4%, respectively. Similarly, the 10th and 90th percentiles (P10 and P90) are -23.2% and 23.6% for Great Plains, and -19.9% and 17.2% for Eastern Temperate Forests. In contrast, the major western ecoregions exhibit much higher variability, with Northwestern Forested Mountains and North American Desert $\sigma$s of 45.6% and 27.0%, respectively, and medians of -6.7% and 4.7%, respectively. P10 and P90 are higher than eastern regions as well, yielding -83.3% and 23.8% for Northwestern Forested Mountains and -33.2% and 36.6% for North American Deserts. Smaller ecoregions (Fig. 8e-j) yield similar spatial trends in variability with the eastern Northern Forests ecoregion showing lower $\sigma$ and P10/90 than the western Marine West Coast Forest and Mediterranean California ecoregions.

The Northwestern Forested Mountains region, accounting for 11% of the CONUS area, is unique among ecoregions in this study. It exhibits the greatest magnitude variability in iterative water budget $R_\text{es}$, P10 - P90 $R_\text{es}$ ranges, skewness (-1.6), and kurtosis (3.7). Furthermore, the difference between the ensemble water budget $R_\text{es}$ and median iterative water budget $R_\text{es}$ is greater than any other ecoregion (-6.3 points).

4 Discussion

Results of model comparisons demonstrate the effect that model disagreement can have on both regional- and continental-scale research. Comparisons of P estimates agree with previous findings (Derin and Yilmaz, 2014; Guirguis and Avissar, 2008; Sun et al., 2018), noting increased uncertainty between products in regions of complex topography. ET comparisons showed that LSMs are more likely to produce lower annual ET than CMs and WBMs, similar to results found by Mueller et al. (2011), and that model disagreement is higher in Pacific regions, similar to the comparison and validation of MOD16-A2, NLDAS2-Noah, and ALEXI by Carter et al. (2018). However, this paper finds some differences from the work of Carter et al. (2018), noting higher uncertainty of ET estimates in arid, semi-arid, and mountainous regions, as well as lower uncertainty in eastern regions. Haddeland et al. (2011) found that LSMs underestimate R partitioning relative to global hydrologic models (not used in this paper). Here, we find that LSMs are more likely to estimate greater R than either WBMs or CMs. Xia et al. (2012b, 2012a) noted greater model disagreement in R estimates between several LSMs in north-eastern and western mountainous regions of the CONUS. This paper, using CV as a measure of relative uncertainty, alternatively shows that modeled runoff uncertainty is higher in the arid and semi-arid regions of the western CONUS where annual R rates are lower.
Previous analyses of modeled SWE estimates focused largely on validation of winter snowpack magnitudes against the SNODAS model or on the timing of accumulation and ablation periods. Broxton et al. (2016) found that LSMs estimated earlier ablation timing than observational measurements, contrasting the results of Essery et al. (2009) and Rutter et al. (2009) who noted later timing in LSMs. Our results indicate the alternative argument that neither snow accumulation nor ablation timing can be accurately attributed solely to model design, noting differences in relative timing of as much as two months between LSMs. Contrasting the arguments by Murdryk et al. (2015) and Broxton et al. (2016), who concluded that uncertainty in estimates may be controlled by differences in model structure, results here show that attributing model differences to structure is likely impossible without controlling for forcing data, model parameters, and calibration methods. Properly identifying controls on model differences requires robust MIPs wherein hydrologic models are operated within the confines of strict input data and calibration schemes, such as those performed by Haddeland et al. (2011), Kollet et al. (2017), or Rosenzweig et al. (2013).

Perhaps the most important variable controlling uncertainties in SM(e) are the discrepancies in model-defined rootzone depth. Typically the largest surface storage component in the hydrologic cycle, differences in SM(e) estimates will more greatly affect water availability calculations than any other component. Koster et al. (2009) argued that modeled soil moisture, related in units of equivalent water depth (i.e. SM(e)), should not be considered direct measures of actual soil water content but instead should be used as relative values to identify seasonal to annual trends and responses to changing climatology. However, LSM estimates of soil moisture depth are commonly applied in research applications as direct measures of soil water content. For example, groundwater storage trends are calculated from Gravity Recovery and Climate Experiment (GRACE) solutions of terrestrial water storage anomalies (Rodell et al., 2007b; Scanlon et al., 2012; Thomas and Famiglietti, 2019; Voss et al., 2013). Groundwater storage is calculated as the remainder after removing surface storage components such as modeled surface water storage, SWE, and SM(e) storage values from GRACE terrestrial water storage. Because estimates of SM(e) can vary by 64-96%, groundwater storage values derived from GRACE solutions will be significantly affected. Groundwater storage values are most commonly extracted using the NLDAS2- or GLDAS-driven LSMs Mosaic, Noah, VIC, or CLM that agree more in monthly SM(e) magnitude than with other estimates from WBM or CMs. However, even those products show differences in SM(e) estimates of up to 250 mm per year. Therefore, quantifying differences in SM(e) is useful in understanding uncertainty propagation in research results.

In terms of evaluating surface water availability, snowmelt from the Northwestern Forested Mountains ecoregion is a primary supply of water for many western population centers. However, disagreement between models of annual snowpack in this study, measured as SWE, averages 97% relative variability and timing on accumulation or ablation can vary by up to two months. These levels of disagreement strongly affect the accuracy of predictive and retrospective snow water analyses of both present snowpack and long-term trends.
LSMs, typically targeted to estimate the vertical ET flux, generally overestimate R compared to CMs. Conversely, CMs, targeted to estimate the horizontal R flux, often overestimate ET compared to LSMs. This is likely an example of equifinality (Beven, 2006), wherein target variable accuracy may be reached without accurately representing the complete hydrologic system. In comparing model output by organization, estimates of P, ET, and SWE produced by NASA are often lower in magnitude than those produced by the USGS and European Centre for Medium-Range Weather Forecasts (ECMWF), indicating organizational differences in model operation methods, such as forcing data selection or calibration methodology.

Results of the water budget case studies demonstrate that water budget Rεs are more variable in the western CONUS, with σ of ±27.0-45.6% by ecoregion, although even eastern CONUS Rεs range in variability from ±14.3-18.9% by ecoregion. In all regions, various combinations of products can result in both positive and negative imbalances (ε), yielding alarmingly different research implications. Positive ε (P > ET + R) means that more water is entering a system than leaving. Assuming limited influence of model uncertainty, this would imply that excluded natural or anthropogenic fluxes are removing water from the P = ET + R system. This could be interpreted as the presence of long-term natural contributions to storage components, such as soil moisture or groundwater, or as the presence of anthropogenic extractions. The same region, depending on models selected, could yield negative ε (P < ET + R), implying the presence of additive fluxes, such as releases from surface storage or increased irrigation from imported water.

Results in this paper supplement the work of Haddeland et al. (2011) who argued that climate change effect modeling on large-scale hydrologic processes should utilize a range of model estimates rather than rely on a single model realization. Indeed, we show that model variability of storage components often exceeds the magnitude of the measurements themselves, calling into question the functional application of terrestrial storage estimates. Our work shows that ensembles of model estimates are not only useful, but are in fact a necessity, to best understand both terrestrial hydrology and quantify our degree of knowledge. Improved ensembling methods, such as was done by Zaherpour et al. (2019), will likely be invaluable resources in the fields of terrestrial hydrology, climatology, and meteorology modeling. Work by Newman et al. (2015) and Addor et al. (2017) organized numerous modeled datasets across hundreds of watershed in the U.S. through the Catchment Attributes and MEteorology for Large-sample Studies (CAMELs) dataset, compiling meteorological, geophysical, and hydrologic response variables. Data hosted within the CAMELs dataset, as well as the datasets prepared for this study (Saxe et al., 2020), reduce the workload for users to better understand the inherent uncertainty in modeled products. This comprehensive and systematic comparison helps elucidate where we as a scientific community could better target research efforts to reconcile large-extent estimates of the hydrologic cycle. Most importantly, differences in modeled SM(e), both in terms of magnitude depth and correlation against RS products, would need to be addressed at the continental-scale to better understand the largest surface store. Further, estimates in SWE, ET, and R need to be reconciled in much of the western CONUS to provide a more accurate retrospective and operational overview of the hydrologic cycle. Finally, differences in hydrologic estimates warrant incorporation into future analysis that use these products to provide the scientific community with a more robust understanding.
of results confidence and areas of uncertainty. All datasets explored in this paper are easily obtained from publicly available sources, and modern data processing workflows are applied to multiple datasets without difficulty.

5 Conclusion

Model selection can significantly alter results and findings when used as a substitute for observational data. Publicly available modeled estimates are commonly used within the scientific community, though literature effectively evaluating differences in magnitude and trend in the context of application is rare. Model Intercomparison Projects abound within the scientific literature, often focusing on the causes of model disagreement (e.g., forcing data, model structure, calibration) and validating against in-situ observational data despite limited availability. In contrast, this study investigated how publicly available modeled estimates of hydrologic flux and storage components can affect scientific analyses by quantifying disagreement between numerous hydrologic models, reanalysis datasets, and remote sensing products.

Results show that flux and storage magnitudes disagree most greatly in the western CONUS, with uncertainty (measured as coefficient of variation) ranging from 11-21% for precipitation (P), 14-26% for evapotranspiration (ET), 28-82% for runoff (R), 76-84% for snow water equivalent (SWE), and 36-96% for soil moisture (SM). In the eastern CONUS, uncertainty is somewhat lower, ranging from 5-14% for P, 13-22% for ET, 28-82% for R, 53-63% for SWE, and 42-83% for SM. Interannual trends in estimates from 1982-2010 show more comprehensive agreement for P and ET fluxes, but common disagreement for R, SWE, and SM. Disagreement in trends (i.e. positive versus negative versus insignificant) between models is typically a result of conflicting negative and insignificant trends rather than between negative and positive trends, indicating that disagreement is due to occurrence and not direction of trend. Correlating fluxes and stores against remote sensing-derived products shows poor overall correlation in the western CONUS for ET and SM. P correlates well in all regions, and SWE correlates well in the primary regions of Northwestern Forested Mountains and Northern Forests.

A water budget analysis, performed by iterating through all combinations of publicly available modeled products highlighted in this research, shows that in large eastern ecoregions model selection can result in relative imbalances ranging from -50 to 50%. In larger western ecoregions, relative imbalances can range from -150 to 60%.

Publicly available modeled estimates of the hydrologic system accelerate the development of scientific research by reducing the necessary workload, processing time, and technical expertise for researchers. In addition, estimates fill knowledge gaps in fluxes and stores where observational data are incomplete. Metrics of disagreement in component estimates presented here help to describe the uncertainty of the scientific community's current state of knowledge. The uncertainty inherent in modern datasets can affect results of studies as diverse as satellite-derived groundwater estimates and predictive snowmelt analyses.

Our results highlight problem areas within CONUS-extent hydrologic estimation efforts that warrant better understanding and addressing in future endeavors. Most importantly, our results show that the most important issues to reconcile are disagreement, in terms of both magnitude and long-term trend, of (a) SM storage across the CONUS, (b) SWE storage in western mountains,
and (c) hydrologic fluxes in western arid and semi-arid regions. Additionally, our results support the findings of previous studies and agree that future work applying modeled data would prove to be more accurate, more informative, and more robust by incorporating model ensembles to provide confidence intervals, better quantify result uncertainty, and improve overall accuracy.

**Acknowledgements:** This work was performed under the U.S. Geological Survey’s Water Availability and Use Science Program and the Hydrologic Sciences and Engineering fellowship from the Colorado School of Mines. The authors are grateful for the early assistance of Flannery Dolan (Tufts University), Shirley Leung (University of Washington), and Adam Price (University of California Santa Cruz). All the data used in this study were obtained from public domains and are freely available. Data used for this study were converted from their primary structures (rasters and shapefiles) to EPA Ecoregions Level 1 and are available in Saxe et al. (2020). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

**References**


Rodell, M., Beaudoin, H.K., NASA/GSFC/HSL, 2007a. GLDAS CLM Land Surface Model L4 3 hourly 1.0 x 1.0 degree Subsetted V001. Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/83NO2QDLG6M0


Running, S., Mu, Q., Zhao, M., 2017. MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC.


Xia, Y. et al., 2012a. NLDAS Noah Land Surface Model L4 Hourly 0.125 x 0.125 degree V002. Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/47Z13FNQODKV

Xia, Y. et al., 2012b. NLDAS Mosaic Land Surface Model L4 Hourly 0.125 x 0.125 degree V002. Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/EN4MBWTCENE5

Xia, Y. et al., 2012c. NLDAS VIC Land Surface Model L4 Hourly 0.125 x 0.125 degree V002. Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/ELBDAPAKNGJ9


Xia, Y., NCEP/EMC, et al., 2009. NLDAS Primary Forcing Data L4 Hourly 0.125 x 0.125 degree V002, Edited by David Mocko. Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/6J5LHHOHZHN4


## Appendix 1

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Product Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hydrologic Models</strong></td>
<td></td>
</tr>
<tr>
<td>CPC</td>
<td><strong>Climate Prediction Center</strong></td>
</tr>
<tr>
<td></td>
<td><em>Rootzone soil moisture (equivalent water depth) estimated through a one-layer water balance model forced by CPC reanalysis precipitation and temperature. Resolutions are monthly at 1/2° from 1948 to present for the globe</em> (Fan &amp; van den Dool, 2004). <em>Author’s acknowledgement: CPC soil moisture data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at <a href="https://www.esrl.noaa.gov/psd/">https://www.esrl.noaa.gov/psd/</a>.</em></td>
</tr>
<tr>
<td>CSIRO-PML</td>
<td><strong>Commonwealth Scientific and Industrial Research Organisation</strong></td>
</tr>
<tr>
<td></td>
<td><em>Estimates of evapotranspiration generated by a spatially explicit Penman-Monteith-Leuning model constrained annually by the Fu hydroclimatic model. The model is driven by precipitation, temperature, vapour pressure, windspeed, and radiation estimates. Datasets used include reanalysis climatology and meteorology products and ex-situ surface and vegetation data. Resolutions are monthly at 1/2° spatial resolution from 1981-2012 for the globe</em> (Zhang et al., 2016).</td>
</tr>
<tr>
<td>ERA5/H-TESSEL</td>
<td><strong>Hydrology revised-Tiled ECMWF Scheme for Surface Exchanges over Land forced with the ERA5 reanalysis</strong></td>
</tr>
<tr>
<td></td>
<td><em>From the European Centre for Medium-Range Weather Forecasts (ECMWF), a land surface model, improved from the original TESSEL model, estimates land surface and subsurface fluxes and stores. The modeled is structured with four soil layers (0-7, 7-28, 28-100, 100-289 cm), a single snow layer, two sub-grid vegetation types, and spatially variant soil type. This paper uses model output forced with ERA5 reanalysis data (see below), provided by C3S in conjunction with the reanalysis product. Resolutions are hourly at 0.25° from 1979 to present for the globe</em> (Balsamo et al., 2009; C3S, 2017). <em>Author’s acknowledgement: Contains modified Copernicus Climate Change Service Information [2019].</em></td>
</tr>
<tr>
<td>ERA5-Land/H-TESSEL</td>
<td><strong>Hydrology revised-Tiled ECMWF Scheme for Surface Exchanges over Land forced with the ERA5-Land reanalysis</strong></td>
</tr>
</tbody>
</table>
From the ECMWF, a land surface model, improved from the original TESSEL model, estimates land surface and subsurface fluxes and stores. See above for a description of model structure. This paper uses model output forced with ERA5-Land reanalysis data (see below), provided by C3S in conjunction with the reanalysis product. Resolutions are hourly at 0.1° from 2001 to present for the globe (Balsamo et al., 2009; C3S, 2019). Author’s acknowledgement: Contains modified Copernicus Climate Change Service Information [2019].

**GLDAS-CLM**

Community Land Model, V2 driven by the NASA GLDAS

From the National Aeronautics and Space Administration (NASA) Global Land Data Assimilation System (GLDAS), a single-column land surface model that estimates surface and subsurface fluxes and stores within independent spatial domains. The model is structured with variable layer spacing for soil temperature and moisture, multilayer snow process parameterization, TOPMODEL-concept runoff parameterization, a canopy photosynthesis-conductance model, and tiling treatments for sub-grid energy and water balances. Model input requirements are land surface types, soil and vegetation parameters, model initialization states, and climatological forcing data. This study uses the operational version driven by the NASA GLDAS. Resolutions are sub-daily at 1° from 1979 to present for the globe (Dai et al., 2003; M. Rodell et al., 2004; Matthew Rodell et al., 2007).

**GLEAM**

Global Land Evaporation Amsterdam Model

Estimates land evaporation, surface soil moisture, rootzone soil moisture, potential evaporation, and evaporative stress conditions. Potential evapotranspiration is calculated using the Priestly-Taylor equation on reanalysis, in-situ, and ex-situ meteorological measurements. Actual evapotranspiration is calculated using an evaporative stress factor based on rootzone soil moisture estimates and microwave ex-situ observations. Rootzone soil moisture is estimated through a water balance model. Model generation 3.3 replaces reanalysis measurements from ERA-Interim with ERA-5 (see below). Model version 3.3b differs from 3.3a by utilizing primarily ex-situ data, excluding reanalysis products. Resolutions are daily at 1/4° from 1980 to present for the globe (Martens et al., 2017; Miralles et al., 2011).

**JRA-25/SiB**

Simple Biosphere model forced with Japanese 25-year Reanalysis

Land surface flux and store estimates generated as output from the Japan Meteorological Agency (JMA)-operated Simple Biosphere model (SiB) forced with Japanese 25-year
ReAnalysis (JRA-25) reanalysis. The predecessor to modern land surface models such as Mosaic, the model is biophysically based and structured with two vegetation layers (canopy and ground cover), and three soil layers. Resolutions are sub-daily to daily at T106 (~110 km) from 1979 to 2004 for the globe (NCAR Research Staff (Eds), 2016; Onogi et al., 2007; Sellers et al., 1986).

**JRA-55/SiB**

Simple Biosphere model forced with Japanese 55-year Reanalysis

Land surface flux and store estimates generated as output from the JMA-operated SiB model forced with Japanese 55-year ReAnalysis (JRA-55) reanalysis. See above for model description. Resolutions are sub-daily to daily at T319 (~55 km) from 1957 to present for the globe (Kobayashi et al., 2015; Kobayashi & NCAR Research Staff (Eds), 2019; Sellers et al., 1986).

**Livneh-VIC**

Variable Infiltration Capacity (VIC) model forced and calibrated by Livneh et al. 2013

Modeled estimates of soil moisture, snow water equivalent, discharge, and surface heat fluxes generated by forcing the VIC model with the Livneh et al. 2013 reanalysis meteorological dataset (see below). See below (NLDAS2-VIC) for description of model structure. Resolutions are sub-daily to daily at 1/16° from 1915 to 2011 for the conterminous United States (Liang et al., 1994; Livneh et al., 2013).

**MERRA-Land/CLSM**

Catchment Land Surface Model forced with MERRA-Land reanalysis

A land surface model developed to improve treatment of horizontal hydrologic process in response to previous land surface model over-attention to vertical processes. The model is structured around tiled hydrologic catchments defined by topography, two soil layers, and three snow layers. This paper uses surface estimates generated from the MERRA-Land-forced version from GEOS-5, released in conjunction with the MERRA-Land reanalysis product (see below). Resolutions are sub-daily at 1/2° from 1980 to 2016 for the globe (Koster et al., 2000; Reichle et al., 2011; Rienecker et al., 2011).

**MERRA-2/CLSM**

Catchment Land Surface Model forced with MERRA-2 reanalysis

A land surface model estimating surface and subsurface hydrologic fluxes and stores. See above for model description. This paper uses estimates generated from the MERRA-2-forced version from GEOS-5, released in conjunction with the MERRA-2 reanalysis product (see
Resolutions are sub-daily at 1/2° from 1980 to present (Gelaro et al., 2017; Reichle et al., 2017).

**NCEP-DOE/Eta-Noah**

Noah land surface model forced with National Centers for Environmental Prediction-Department of Energy, R-2 reanalysis

Noah land surface model estimates surface and subsurface hydrologic fluxes. See below (NLDAS2-Noah) for model description. Model estimates here are generated through the Noah land surface model component of the NCEP Eta atmospheric model forced with NCEP-DOE reanalysis (see below), released in conjunction with the NCEP-DOE product. Resolutions are sub-daily at T62 (~210 km) gridding from 1979 to present for the globe (Kalnay et al., 1996; Kanamitsu et al., 2002). Author acknowledgment: NCEP Reanalysis 2 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/.

**NCEP-NARR/Eta-Noah**

Noah land surface model forced with National Centers for Environmental Prediction-North American Regional Reanalysis

Noah land surface model estimates produced as a component of the larger Eta atmospheric model. See below (NLDAS2-Noah) for model description. Model estimates here are generated through the Noah surface model component of the NCEP Eta atmospheric model forced with NCEP North American Regional Reanalysis (NCEP-NARR), released in conjunction with the NCEP-NARR product. Resolutions are sub-daily at 32 km from 1979 to present for North America (Mesinger et al., 2006). Author acknowledgement: NCEP Reanalysis data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/.

**NHM-MWBM**

National Hydrologic Model-framework Monthly Water Balance Model

A water balance model applied within the U.S. Geological Survey’s National Hydrologic Model framework that utilizes a monthly accounting procedure to estimate evapotranspiration, runoff, soil moisture, and snow water equivalent based on methodology from Thornthwaite (1948). Resolutions are monthly at Hydrologic Response Units from 1949 to 2010 for the conterminous United States (G. J. McCabe & Markstrom, 2007).

**NHM-PRMS**

National Hydrologic Model-framework Precipitation Runoff Modeling System
A process-based, deterministic hydrologic model applied within the framework of the U.S. Geological Survey’s National Hydrologic Model that estimates various surface and subsurface fluxes and stores using a conceptualized watershed composed of a series of reservoirs, stream segments, and lakes maintained with a balanced water budget. The model utilizes reanalysis and ex-situ physical characteristic data of topography, soils, vegetation, geology, and land use to derive required parameters, and is driven by precipitation and temperature reanalysis products. Resolutions are daily at Hydrologic Response Units from 1980 to 2016 for the conterminous United States (Markstrom et al., 2015; Regan et al., 2018).

**NLDAS2-Mosaic**

Mosaic model driven by the NASA NLDAS, Phase 2

A land surface model, directly descended from the SiB land surface model, that estimates surface and subsurface fluxes and stores, originally targeted to be coupled with climate and weather models. The model is structured to allow vegetation control over surface energy and water balances, a canopy interception reservoir, three soil reservoirs (thin surface, middle “rootzone”, and lower “recharge”), and a complete snow budget. The model name is derived from the mosaic approach of tiling sub-grid cells into homogeneous vegetation types with independent energy balances. This study uses the operational version driven by the NASA NLDAS Phase 2 (see below), wherein the model was configured to support up to 10 tiles per grid cell, each with specified predominant soil types and three spatially invariant soil layers. Resolutions are hourly at 1/8° from 1979 to present for North America (Koster & Suarez, 1996; Y. Xia et al., 2012a; Youlong Xia, Mitchell, Ek, Sheffield, et al., 2012b).

**NLDAS2-Noah**

Noah model driven by the NASA NLDAS, Phase 2

A grid-based land surface model, simpler than the Noah-MP version, originally developed as a component of the NOAA-NCEP Eta model and updated for use in the NASA NLDAS, Phase 2 (see below), and is used within the Weather Research and Forecasting (WRF) atmospheric model, the National Oceanic and Atmospheric Administration-NCEP Climate Forecast System (NOAA-NCEP CFS), and the NOAA Global Forecast System. The model is structured with four spatially invariant thickness soil layers (three layers forming the rootzone system in non-forested regions and four in forested regions) and can simulate soil freeze-thaw processes. This study uses the operational version (Noah-2.8) driven by the NASA NLDAS, Phase 2 (see below). Resolutions are hourly at 1/8° from 1979 to present for North America (Ek et al., 2003; Y. Xia et al., 2012b; Youlong Xia, Mitchell, Ek, Sheffield, et al., 2012b).
NLDAS2-VIC  Variable Infiltration Capacity model driven by the NLDAS, Phase 2  
A semi-distributed, grid-based hydrologic model estimating surface and subsurface fluxes and stores. Model structure is composed of three soil layers (top spatially invariant 10 cm thickness, others spatially-variant) with rootzone depth controlled by vegetation. The model utilizes sub-grid vegetation tiling similar to the Mosaic land surface model (see above), and a two-layer energy balance snow model. This study uses the operational version driven by the NASA NLDAS, Phase 2 (see below). Resolutions are hourly at 1/8° from 1979 to present for North America (Liang et al., 1994; Y. Xia et al., 2012c; Youlong Xia, Mitchell, Ek, Sheffield, et al., 2012b).

SNODAS  SNOw Data Assimilation System  
Estimates snow cover and snow water equivalent by integrating modeled snow estimates with observational and reanalysis data from in-situ and ex-situ sources. The primary estimation method is a physically based, spatially-distributed energy- and mass-balance snow model forced by downscaled output from the National Weather Service Rapid Refresh weather forecast model. In-situ and ex-situ data sources are applied depending on difference fields between modeled and observed values, and used to perform immediate model calibrations. Resolutions are daily at 1 km from 2003 to present for North America (Barrett, 2003; National Operational Hydrologic Remote Sensing Center, 2004).

TerraClimate  TerraClimate  
A dataset utilizing both reanalysis and water budget modeling to generate estimates of climate, meteorological, and hydrologic variables. Climate and meteorological estimates are calculated by applying interpolated time-varying anomalies from both CRU Ts4.0 and JRA-55 to the higher resolution WorldClim climatology. Hydrologic estimates of evapotranspiration, precipitation, temperature, and soil water capacity are generated through a modified Thornthwaite-Mather climatic water balance model. Water balance-derived resolutions are monthly at 1/2° from 1958 to present for the globe (Abatzoglou et al., 2018).

VegET  Vegetation ET  
Estimates rootzone soil moisture and evapotranspiration through a precipitation-driven one-dimensional rootzone water balance model. Because the model is parameterized to operate on
a control volume using water holding capacity, it captures only evapotranspiration from vegetation sources (i.e. natural conditions) and does not consider anthropogenic (i.e. non-natural) water use. Data provided via personal correspondence with the authors. Resolutions are daily at 1 km from 2000 to 2014 for the conterminous United States (Gabriel B. Senay, 2008; Naga Manohar Velpuri & Senay, 2017).

<table>
<thead>
<tr>
<th>Reanalysis</th>
<th>Canadian Sea Ice and Snow Evolution Network, V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates snow water equivalent (SWE) by merging five observation-based estimates through an adapted ensemble mean methodology. Merged products are (1) GlobSnow-combined SWE (merges ex-situ passive microwave and in-situ weather station observations), (2) ERA-Interim reanalysis, (3) MERRA reanalysis, (4) SWE from the Interaction Sol-Biosphère-Atmosphère (ISBA) land surface model forced with ERA-Interim reanalysis (see below), and (5) NASA Global Land Data Assimilation System reanalysis. Resolutions are daily at 1° from 1981 to 2010 for the Northern Hemisphere (Mudryk et al., 2015; Mudryk &amp; Derksen, 2017).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CMAP</th>
<th>CPC Merged Analysis of Precipitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>From the CPC, estimates precipitation by blending various in-situ, ex-situ, and reanalysis datasets. The “Standard” product estimates precipitation by blending ex-situ Global Precipitation Index (GPI), OLR Precipitation Index (OPI), Special Sensor Microwave/Imager (SSM/I) emission, and Microwave Sounding Unit (MSU) data with in-situ precipitation gauge measurement data. The “Enhanced” product additionally includes blended NCEP-NCAR reanalysis (see below) precipitation estimates. Resolutions are monthly at 2 1/2° from 1979 to present for the globe (Xie &amp; Arkin, 1997). Author’s acknowledgement: CMAP Precipitation data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at <a href="https://www.esrl.noaa.gov/psd/">https://www.esrl.noaa.gov/psd/</a>.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DayMET</th>
<th>Daily Surface Weather Data on a 1-km Grid for North America, V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 3 of the DayMET model, generates estimates of temperature, precipitation, radiation, vapor pressure, snow water equivalent, and day length from in-situ observations of temperature and precipitation and digital elevation data. Interpolation is performed using the spatial convolution of a truncated Gaussian weighting filter with an iterative station density algorithm</td>
<td></td>
</tr>
</tbody>
</table>
applied to heterogeneous observation distribution in complex terrain. Snow water equivalent is estimated using a simple temperature-based snowmelt model from Running and Coughlan (1988). This paper assigns DayMET snow water equivalent estimates to the “Reanalysis” category because of the snow model’s simplicity and lack of a physical or conceptual hydrologic framework. Descriptions of the current model framework are best described at the ORNL DAAC website (daac.ornl.gov/DAYMET/guides/Daymet_V3_CF_Mosaics.html). Resolutions are daily at 1 km from 1980 to present for North America (P.E. Thornton et al., 2017; Peter E. Thornton et al., 1997, 2000).

ERA5

ECMWF ReAnalysis, Fifth Product

Generates numerous estimates of climate and meteorological variables by assimilating observational data from 55 ex-situ and 19 in-situ sources using the 4D-Var variational method. Previous generations of the ERA product were FGGE, ERA-15, ERA-40, and ERA-Interim, all now out of service. Resolutions are hourly at 0.25° from 1979 to present for the globe (C3S, 2017). Author’s acknowledgement: Contains modified Copernicus Climate Change Service Information [2019].

ERA5-Land

ECMWF Reanalysis, Fifth Product

Estimates of climate, meteorological, and surface hydrology fluxes by replaying the ERA5 reanalysis (see above) land component at higher spatial resolution. Resolutions are hourly at 0.1° from 2001 to present for the globe (Balsamo et al., 2009; C3S, 2019). Author’s acknowledgement: Contains modified Copernicus Climate Change Service Information [2019].

GPCC

Global Precipitation Climatology Centre

Monthly accumulated precipitation estimates based on 67,200 in-situ observational stations with decade or longer temporal spans. The dataset used in this paper is V7, one of two GPCC Full Data Products, which the product’s authors suggest are the highest accuracy of their datasets. Resolutions are monthly at 1/2° from 1901 to 2013 for the globe (Becker et al., 2013; Schneider et al., 2011). Author’s acknowledgement: GPCC Precipitation data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/

gridMET

gridMET (or, METDATA)
Estimates temperature, precipitation, downward shortwave radiation, wind-velocity, humidity, relative humidity, and specific humidity by blending PRISM reanalysis (see below) climate data with NLDAS2 reanalysis (see below) data. Resolutions are daily at 1/24° from 1979 to present for the conterminous United States (Abatzoglou, 2013).

<table>
<thead>
<tr>
<th>Livneh et al. 2013</th>
<th>Livneh daily CONUS near-surface gridded observed meteorological data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Near-surface meteorological estimates generated from approximately 20,000 NOAA Cooperative Observer station daily datasets. Precipitation and temperature in-situ observations are converted to grids using the synergraphic mapping system, wind data linearly interpolated from the lower resolution NCEP-NCAR reanalysis dataset. Other variables were derived using methods from the mountain microclimate simulator. Daily temperature data is converted to 3-hourly using spline interpolation. Resolutions are daily at 1/16° from 1915 to 2011 for the conterminous United States (Livneh et al., 2013). Author acknowledgement: Livneh data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at <a href="https://www.esrl.noaa.gov/psd/">https://www.esrl.noaa.gov/psd/</a>.</td>
</tr>
</tbody>
</table>

| Maurer et al 2002 | Meteorological and surface hydrologic estimates generated through both reanalysis and hydrologic modeling. Precipitation gridded from in-situ daily measurement stations using the synergraphic mapping system algorithm scaled to match PRISM (see below) long-term averages. Surface estimates derived through the VIC (see above) hydrologic model. Resolutions are daily at 1/8° from 1950-1999 for North America (Maurer et al., 2002). |

<table>
<thead>
<tr>
<th>MERRA-Land</th>
<th>Modern-Era Retrospective analysis for Research and Applications with improved land surface variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The predecessor to MERRA-2 (see below) estimates meteorological and surface hydrologic components. Meteorological estimates are derived through the GEOS-5 atmospheric general circulation model by assimilating in-situ (3DVAR analysis algorithm) and ex-situ (Community Radiative Transfer Model) observational data. This paper uses the Land Surface Diagnostics precipitation. Resolutions are sub-daily at 1/2° from 1980 to 2016 for the globe (Rienecker et al., 2011).</td>
</tr>
</tbody>
</table>

| MERRA-2 | Modern-Era Retrospective analysis for Research and Applications, V2 |
The most recent version of the MERRA product differs from MERRA-land by including updates to models, algorithms, observing systems, and ex-situ processing methods, as well as 14 additional ex-situ data sources. Precipitation is derived from global precipitation products disaggregated to hourly time steps using MERRA-Land precipitation. Precipitation estimates used in this paper are from the Land Surface Diagnostics category. Resolutions are sub-daily at 1/2° from 1980 to present (Gelaro et al., 2017).

NCEP-DOE
National Centers for Environmental Prediction-Department of Energy, R-2
The updated (R-2) version of the NCEP-NCAR reanalysis, estimates climate, meteorological, and surface hydrologic components. Reanalysis estimates of atmospheric variables are performed through assimilation and modeling of in-situ and ex-situ observational data. Resolutions are sub-daily at T62 (~210 km) gridding from 1979 to present for the globe (Kalnay et al., 1996; Kanamitsu et al., 2002). Author acknowledgment: NCEP-DOE data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/.

NLDAS2
North American Land Data Assimilation System, Phase 2
Compiled data used to drive land surface models (see above), including (1) land surface parameters of vegetation, soil, topography, and temperature; (2) surface forcing fields of precipitation, radiation, temperature, humidity, wind, and pressure derived by blending various in-situ and ex-situ datasets. Precipitation, specifically, is derived by blending temporally disaggregated reanalysis (CPC and NARR, see above) and ex-situ (Doppler Stage II and CMORPH) estimates. Resolutions are hourly at 1/8° from 1979 to present for North America (Y. Xia, NCEP/EMC, et al., 2009; Youlong Xia, Mitchell, Ek, Sheffield, et al., 2012b).

PRISM
Parameter-elevation Regressions on Independent Slopes Model
Reanalysis estimates of precipitation and temperature calculated using a climate-elevation regression model, utilizing information of location, elevation, coastal proximity, and other geophysical parameters. Resolutions are daily at 800 m (paid) and 4 km (free) from 1895 to present for the United States (PRISM Climate Group, OSU, 2004).

Reitz et al 2017
Reitz et al. 2017
Reanalysis estimates of groundwater recharge, quick-flow runoff, and evapotranspiration. Runoff and evapotranspiration are calculated using regression equations derived from observational water balance data using land cover, temperature, and precipitation information. Groundwater recharge is calculated as the remainder of a water balance using runoff, evapotranspiration, and precipitation. Resolutions are annual (monthly provided via personal correspondence with authors) at 800 m from 2000 to 2013 for the conterminous United States (Reitz et al., 2017).

UoD-v5

University of Delaware Air Temperature and Precipitation, V5
Reanalysis precipitation and temperature estimates calculated from in-situ observational data with an interpolation algorithm based on the spherical Shepard’s distance-weighting method with Digital Elevation Model information. This paper uses the fifth product version. Resolutions are monthly at 1/2° from 1950 to 1999 (Willmott & Matsuura, 2001). Author acknowledgement: UoD precipitation data provided by the NOAA/ OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/.

WaterWatch

Estimates of runoff derived from the U.S. Geological Survey’s in-situ streamgauge network. Resolutions are monthly at watershed, Hydrologic Unit Levels 2-8, and state polygons from 1901 to present (Jian et al., 2008).

Remote Sensing

AMSR-E/Aqua
AMSR-E/Aqua Monthly L3 Global Snow Water Equivalent EASE-Grids, V2
Scientifically identical to Version 1 (updates to product maturity code in V2), estimates SWE using passive microwave data collected from the Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) instrument hosted on NASA’s Aqua satellite. Microwave measurements are converted to SWE estimates using the AMSR-E Snow Water Equivalent Algorithm that utilizes brightness temperature differences calculated using the dense media radiative transfer equation and pre-selected snowpack profiles to develop an artificial neural network. Probable SWE estimate ranges are then restricted using surface temperature and land cover attributes derived from the ex-situ MODIS sensor hosted on the NASA Terra
satellite. Resolutions are daily at 25 km from 2002 to 2011 for the globe (A. T. C. Chang et al., 2003; Alfred T.C. Chang & Rango, 2000; Tedesco et al., 2004).

**ESA-CCI**

European Space Agency – Climate Change Initiative

Volumetric soil moisture estimates generated by merging both active and passive ex-situ microwave data-derived soil moisture products into three products: ACTIVE, PASSIVE, and COMBINED. Passive microwave products are from the Scanning Multichannel Microwave Radiometer (SMMR), SSM/I, Tropical Microwave Imager (TMI), Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E), WindSat, Soil Moisture and Ocean Salinity (SMOS), and AMSR2 sensors. Active microwave products are from the Active Microwave Instrument Wind Scatterometer (AMI-WS), Advanced SCATterometer – satellite A or satellites A/B (ASCAT-A and ASCAT-A/B) sensors. Merging is performed by resampling target products to a uniform spatiotemporal structure, scaling to match ranges, and weighted using triple collocation. This research utilizes the COMBINED product. Resolutions are daily at 1/4° from 1978 to present for the globe (Dorigo et al., 2017).

**GPCP-v3**

Global Precipitation Climatology Project, V3 Precipitation Data (Beta)

Estimates of precipitation generated by assimilating radiometer data from SSMI and SSMIS, infrared data from the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record (PERSIANN-CDR), TIROS Operational Vertical Sounder (TIROS-TOVS) sensor and Tropical Rainfall Measuring Missing (TRMM) Combined Climatology precipitation estimates with the GPCC global gauge analysis (see above in “Reanalysis”). This paper uses the “combined satellite-gauge” product rather than the “satellite-only” product. Resolutions are monthly at 1/2° from 1983 to 2016 for the globe (G. J. Huffman et al., 2019).

**MOD16-A2**

MODIS Global Evapotranspiration Project

Level 4 Moderate Resolution Imaging Spectroradiometer (MODIS) land data product that estimates evapotranspiration from ex-situ observational MODIS vegetation data and reanalysis meteorological data using an improved evapotranspiration model (Mu et al., 2011) based on the Penman-Monteith equation. Resolutions are 8-day at 1 km from 2000-2010 for the globe (S. Running et al., 2017).
<table>
<thead>
<tr>
<th><strong>SMOS-L4</strong></th>
<th>Soil Moisture and Ocean Salinity - Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Volumetric rootzone soil moisture estimates derived using CATDS-generated ascending and descending orbit SMOS L3 surface volumetric soil moisture estimates in conjunction with MODIS sensor-derived vegetation information, reanalysis NCEP climate data, FAO soil textures, and ECOCLIMAP surface cover. A double bucket model, composed of a simple water budget model (5-40 cm depth) and a budget model based on a linearized Richard’s Equation formulation (40-200 cm depth), is used to extrapolate surface SMOS L3 soil moisture (0-5 cm) to the rootzone domain. This study merged the ascending and descending orbit products by mean. Resolutions are daily at 25 km EASE Grids from 2010 to 2017 for the globe (Al Bitar et al., 2013).</em>**</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>SSEBop</strong></th>
<th>Operational Simplified Surface Energy Balance model</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A parameterization of the Simplified Surface Energy Balance approach that estimates evapotranspiration from ex-situ MODIS vegetation, ex-situ Shuttle Radar Topography Mission (SRTM) elevation data, reanalysis meteorological data, and other modeled data using an energy balance approach where actual evapotranspiration is calculated as the difference between net surface radiation, sensible heat flux, and ground heat flux. Resolutions are monthly at 1 km from 2000 to 2014 for the conterminous United States and portions of surrounding countries (Gabriel B. Senay et al., 2013; G.B. Senay et al., 2011).</em></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>TMPA-3B43</strong></th>
<th>TRMM Multisatellite Precipitation Analysis, V7</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>From the larger Tropical Rainfall Measuring Mission (TRMM), generates spatiotemporally continuous precipitation estimates by merging precipitation estimates from numerous ex-situ data sources. Primary data sources include (1) passive microwave data from various low earth orbit satellites (e.g. TRMM, AMSR-E) that are converted to precipitation estimates using source-specific algorithms; (2) infrared data collected by geosynchronous earth orbit satellites; (3) TRMM Combined Instrument estimate, a merged passive microwave and active radar product; 4) in-situ GPCP and CAMS monthly precipitation measurements. Resolutions are sub-daily at 1/4° from 1998 to present for the globe (George J. Huffman et al., 2007, 2010).</em></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 2

Figures A2.1 – 2.10 (see below)
### Tables

**Table 1:** Summary of datasets used in this research including assigned data category, abbreviated name, primary organization, literature reference, spatiotemporal resolution, and sourced hydrologic components. The reader is referred to Appendix 1 for definitions of abbreviated model names, as well as model descriptions and further references. Components are precipitation (P), evapotranspiration (ET), runoff (R), snow water equivalent (SWE), and soil moisture in equivalent water depth and volumetric water content (SM(e) & SM(v), respectively). Hydrologic models NMH-MWBM and -PRMS are based on a delineated (i.e. non-gridded), topographically derived spatial framework composed on Hydrologic Response Units (HRUs). The reanalysis product WaterWatch is generated at Hydrologic Units (HU) 2-8. The finest resolution product, HU8, is used in this study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Group</th>
<th>Reference</th>
<th>Spatiotemporal</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hydrologic Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPC</td>
<td>CPC</td>
<td>(Fan &amp; van den Dool, 2004)</td>
<td>$1/2^\circ$</td>
<td>1948-present</td>
</tr>
<tr>
<td>CSIRO-PML‡</td>
<td>CSIRO</td>
<td>(Zhang et al., 2016)</td>
<td>$1/2^\circ$</td>
<td>1981-2012</td>
</tr>
<tr>
<td>ERA5/H-TESSEL‡</td>
<td>ECMWF</td>
<td>(C3S, 2017)</td>
<td>$1/4^\circ$</td>
<td>1979-present</td>
</tr>
<tr>
<td>ERA5-Land/H-TESSEL</td>
<td>ECMWF</td>
<td>(C3S, 2017)</td>
<td>$1/10^\circ$</td>
<td>2001-present</td>
</tr>
<tr>
<td>GLDAS-CLM†</td>
<td>NASA</td>
<td>(Matthew Rodell et al., 2007)</td>
<td>$1^\circ$</td>
<td>1979-present</td>
</tr>
<tr>
<td>GLEAM*,‡</td>
<td>U. of BE</td>
<td>(Martens et al., 2017)</td>
<td>$1/4^\circ$</td>
<td>1980-present</td>
</tr>
<tr>
<td>JRA-25/SiB</td>
<td>JMA</td>
<td>(Onogi et al., 2007)</td>
<td>110km</td>
<td>1979-present</td>
</tr>
<tr>
<td>JRA-55/SiB‡</td>
<td>JMA</td>
<td>(Kobayashi et al., 2015)</td>
<td>55km</td>
<td>1957-present</td>
</tr>
<tr>
<td>Livneh-VIC</td>
<td>CIRES</td>
<td>(Livneh et al., 2013)</td>
<td>$1/16^\circ$</td>
<td>1915-2011</td>
</tr>
<tr>
<td>MERRA-Land/CLSM‡</td>
<td>NASA</td>
<td>(Reichle et al., 2011)</td>
<td>$1/2^\circ$</td>
<td>1980-2016</td>
</tr>
<tr>
<td>MERRA-2/CLSM‡</td>
<td>NASA</td>
<td>(Reichle et al., 2017)</td>
<td>$1/2^\circ$</td>
<td>1980-present</td>
</tr>
<tr>
<td>NCEP-DOE/Eta-Noah†</td>
<td>NOAA</td>
<td>(Kanamitsu et al., 2002)</td>
<td>210 km</td>
<td>1979-present</td>
</tr>
<tr>
<td>NCEP-NARR/Eta-Noah</td>
<td>NOAA</td>
<td>(Mesinger et al., 2006)</td>
<td>32 km</td>
<td>1979-present</td>
</tr>
<tr>
<td>NHM-MWBM‡</td>
<td>USGS</td>
<td>(G. J. McCabe &amp; Markstrom, 2007)</td>
<td>HRU</td>
<td>1949-2010</td>
</tr>
<tr>
<td>NHM-PRMS‡</td>
<td>USGS</td>
<td>(Regan et al., 2018)</td>
<td>HRU</td>
<td>1980-2016</td>
</tr>
<tr>
<td>NLDAS2-Mosaic‡</td>
<td>NASA</td>
<td>(Y. Xia et al., 2012a)</td>
<td>$1/8^\circ$</td>
<td>1979-present</td>
</tr>
<tr>
<td>NLDAS2-Noah‡</td>
<td>NASA</td>
<td>(Y. Xia et al., 2012b)</td>
<td>$1/8^\circ$</td>
<td>1979-present</td>
</tr>
<tr>
<td>NLDAS2-VIC†</td>
<td>NASA</td>
<td>(Y. Xia et al., 2012c)</td>
<td>$1/8^\circ$</td>
<td>1979-present</td>
</tr>
<tr>
<td>SNODAS</td>
<td>NWS</td>
<td>(Barrett, 2003)</td>
<td>1 km</td>
<td>2003-present</td>
</tr>
<tr>
<td>TerraClimate‡</td>
<td>U. of ID</td>
<td>(Abatzoglou et al., 2018)</td>
<td>$1/24^\circ$</td>
<td>1958-present</td>
</tr>
<tr>
<td>VegET†</td>
<td>USGS</td>
<td>(Gabriel B. Senay, 2008)</td>
<td>1 km</td>
<td>2000-2014</td>
</tr>
<tr>
<td><strong>Reanalysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CanSISE</td>
<td>U. Toronto</td>
<td>(Mudryk &amp; Derksen, 2017)</td>
<td>$1^\circ$</td>
<td>1981-2010</td>
</tr>
<tr>
<td>CMAP‡,†</td>
<td>CPC</td>
<td>(Xie &amp; Arkin, 1997)</td>
<td>2 ¾°</td>
<td>1979-present</td>
</tr>
<tr>
<td>DayMET‡</td>
<td>ORNL</td>
<td>(M. . Thornton et al., 2018)</td>
<td>1 km</td>
<td>1980-present</td>
</tr>
<tr>
<td>ERA5‡</td>
<td>ECMWF</td>
<td>(C3S, 2017)</td>
<td>$1/4^\circ$</td>
<td>1979-present</td>
</tr>
<tr>
<td>ERA5-Land</td>
<td>ECMWF</td>
<td>(C3S, 2019)</td>
<td>$1/10^\circ$</td>
<td>2001-present</td>
</tr>
<tr>
<td>GPCC‡</td>
<td>GPCC</td>
<td>(Becker et al., 2013)</td>
<td>$1/2^\circ$</td>
<td>1901-2013</td>
</tr>
<tr>
<td>gridMET‡</td>
<td>U. of ID</td>
<td>(Abatzoglou, 2013)</td>
<td>$1/24^\circ$</td>
<td>1979-present</td>
</tr>
</tbody>
</table>
Livneh et al. 2013‡
Maurer et al. 2002
MERRA-Land‡
MERRA-2‡
NCEP-DOE‡
NLDAS2‡
PRISM‡
Reitz et al. 2017‡
UoD-v5‡
WaterWatch‡
Remote Sensing
AMSR-E/Aqua
ESA-CCI
GPCP-v3‡
MOD16-A2‡
SMOS-L4
SSEBop‡
TMPA-3B43‡

Livneh et al. (Livneh et al., 2013)‡
Maurer et al. (Maurer et al., 2002)‡
Rienecker et al. (Rienecker et al., 2011)‡
Gelaro et al. (Gelaro et al., 2017)‡
Kanamitsu et al. (Kanamitsu et al., 2002)‡
Y. Xia, NCEP/EMC, et al. (Y. Xia, NCEP/EMC, et al., 2009)‡
PRISM Climate Group (PRISM Climate Group, 2004)‡
Willmott & Matsuura (Willmott & Matsuura, 2001)‡
WaterWatch (Jian et al., 2008)‡
Tedesco et al. (Tedesco et al., 2004)‡
Dorigo et al. (Dorigo et al., 2017)‡
G. J. Huffman et al. (G. J. Huffman et al., 2019)‡
S. Running et al. (S. Running et al., 2013)‡
George J. Huffman et al. (George J. Huffman et al., 2010)‡

版本信息
- Versions 3.3a & 3.3b
- Versions Standard & Enhanced
- Datasets used in the water budget case study with complete data for water years 2001-2010

### Table 1: Sizes of study ecoregions in square kilometres [km²] and percent of study domain.

<table>
<thead>
<tr>
<th>Ecoregion</th>
<th>Area [km²]</th>
<th>Percent of CONUS [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marine West Coast Forest</td>
<td>91,035</td>
<td>1.11</td>
</tr>
<tr>
<td>Mediterranean California</td>
<td>168,757</td>
<td>2.06</td>
</tr>
<tr>
<td>Northwestern Forested Mountains*</td>
<td>859,801</td>
<td>10.52</td>
</tr>
<tr>
<td>North American Deserts*</td>
<td>1,533,536</td>
<td>18.76</td>
</tr>
<tr>
<td>Southern Semiarid Highlands</td>
<td>66,699</td>
<td>0.82</td>
</tr>
<tr>
<td>Temperate Sierras</td>
<td>110,038</td>
<td>1.35</td>
</tr>
<tr>
<td>Great Plains*</td>
<td>2,328,673</td>
<td>28.48</td>
</tr>
<tr>
<td>Eastern Temperate Forests*</td>
<td>2,561,292</td>
<td>31.33</td>
</tr>
<tr>
<td>Northern Forests</td>
<td>433,032</td>
<td>5.30</td>
</tr>
<tr>
<td>Tropical Wet Forests</td>
<td>22,469</td>
<td>0.27</td>
</tr>
<tr>
<td><strong>Total Domain Area</strong></td>
<td><strong>8,175,332</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

*Primary ecoregions highlighted during water budget analysis
Table 3: Spearman’s rho ($\rho$) correlation of monthly hydrologic model and reanalysis dataset estimates against remote sensing products quantifying hydrologic components precipitation (P), evapotranspiration (ET), snow water equivalent (SWE), and soil moisture (SM). Correlation was calculated at ten ecoregions and the CONUS extent. The number of modelled datasets correlated against remote sensing products is shown as ($n = X$) for each column. Mean and standard deviation ($\sigma$) of $\rho$ are provided for each ecoregion and remote sensing products as $\rho \pm \sigma$. Counts of statistically insignificant correlations ($Y$) are denoted when $n > 0$ as ($Y$). SM combines datasets produced in units of equivalent water depth ($SM(e)$ in text) and volumetric soil moisture water content ($SM(v)$ in text). Correlation of SWE products is included only for regions of high annual mean SWE storage. Hydrologic model and reanalysis datasets (Table 1) used in each column are listed in the footnotes.

<table>
<thead>
<tr>
<th>Ecoregion</th>
<th>P</th>
<th>AET</th>
<th>SWE</th>
<th>RZSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GPCP-v3$^1$ ($n=15$)</td>
<td>TMUSA-3843$^2$ ($n=14$)</td>
<td>MOD16-A2$^3$ ($n=15$)</td>
<td>SSEbop$^4$ ($n=15$)</td>
</tr>
<tr>
<td>Marine W. Coast Forest</td>
<td>$0.99 \pm 0.01$</td>
<td>$0.99 \pm 0.01$</td>
<td>$0.82 \pm 0.18$</td>
<td>$0.80 \pm 0.19$</td>
</tr>
<tr>
<td>Medit. California</td>
<td>$0.98 \pm 0.01$</td>
<td>$0.97 \pm 0.02$</td>
<td>$0.71 \pm 0.20$</td>
<td>$0.25 \pm 0.42$ (4)</td>
</tr>
<tr>
<td>NW Forested Mtns</td>
<td>$0.97 \pm 0.05$</td>
<td>$0.96 \pm 0.05$</td>
<td>$0.91 \pm 0.05$</td>
<td>$0.89 \pm 0.08$</td>
</tr>
<tr>
<td>N. American Deserts</td>
<td>$0.95 \pm 0.05$</td>
<td>$0.94 \pm 0.05$</td>
<td>$-0.28 \pm 0.17$ (4)</td>
<td>$0.81 \pm 0.13$</td>
</tr>
<tr>
<td>So. Semiarid Highlands</td>
<td>$0.97 \pm 0.04$</td>
<td>$0.96 \pm 0.05$</td>
<td>$0.67 \pm 0.17$</td>
<td>$0.64 \pm 0.17$</td>
</tr>
<tr>
<td>Temperate Sierras</td>
<td>$0.97 \pm 0.04$</td>
<td>$0.96 \pm 0.04$</td>
<td>$0.59 \pm 0.15$</td>
<td>$0.68 \pm 0.18$</td>
</tr>
<tr>
<td>Great Plains</td>
<td>$0.99 \pm 0.02$</td>
<td>$0.98 \pm 0.01$</td>
<td>$0.92 \pm 0.02$</td>
<td>$0.93 \pm 0.03$</td>
</tr>
<tr>
<td>E. Temperate Forests</td>
<td>$0.95 \pm 0.02$</td>
<td>$0.94 \pm 0.07$</td>
<td>$0.96 \pm 0.03$</td>
<td>$0.94 \pm 0.03$</td>
</tr>
<tr>
<td>Northern Forests</td>
<td>$0.96 \pm 0.03$</td>
<td>$0.93 \pm 0.02$</td>
<td>$0.94 \pm 0.03$</td>
<td>$0.94 \pm 0.02$</td>
</tr>
<tr>
<td>Tropical Wet Forests</td>
<td>$0.97 \pm 0.03$</td>
<td>$0.96 \pm 0.03$</td>
<td>$0.88 \pm 0.05$</td>
<td>$0.88 \pm 0.09$</td>
</tr>
<tr>
<td>CONUS$^*$</td>
<td>$0.94 \pm 0.07$</td>
<td>$0.91 \pm 0.06$</td>
<td>$0.95 \pm 0.02$</td>
<td>$0.95 \pm 0.03$</td>
</tr>
</tbody>
</table>

$^1$ CMORPH-enhanced & Standard: DayMET, ERA5, ERA-Land, GPCP, NCEP, and GLEAM
$^2$ CMORPH-enhanced & Standard: CMORPH, ERA5, ERA-Land, GPCP, and GLEAM
$^3$ CMORPH-enhanced & Standard: DayMET, ERA5, ERA-Land, GPCP, and GLEAM
$^4$ CMORPH-enhanced & Standard: DayMET, ERA5, ERA-Land, GPCP, and GLEAM
$^5$ CMORPH-enhanced & Standard: DayMET, ERA5, ERA-Land, GPCP, and GLEAM
$^6$ CMORPH-enhanced & Standard: DayMET, ERA5, ERA-Land, GPCP, and GLEAM

Regions omitted.

* Calculated from datasets spatially aggregated by weighted area mean at the CONUS extent (Figure 2), not as a mean of individual ecoregion estimates above.
Table 4: Summary statistics of ecoregion water budget calculations summed for complete water years 2001-2010. Median ($P_{50}$) and standard deviation ($\sigma$) summarize the iterative water budgets’ ($n = 2,925$) relative imbalances ($Re$) in units of percent. The $10^{th}$, $25^{th}$, $75^{th}$, and $90^{th}$ percentiles ($P_i$) quantify distribution ranges, along with skew ($\gamma$) and kurtosis ($K$). The ensemble column $\mu$ provides water budget $Re$ produced from the ensemble mean of all datasets available for water years 2001-2010.

<table>
<thead>
<tr>
<th>Ecoregion</th>
<th>$\sigma$</th>
<th>$P_{10}$</th>
<th>$P_{25}$</th>
<th>$P_{50}$</th>
<th>$P_{75}$</th>
<th>$P_{90}$</th>
<th>$\mu$</th>
<th>$\gamma$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marine W. Coast Forest</td>
<td>±29.5</td>
<td>-46.0</td>
<td>-21.3</td>
<td>-2.3</td>
<td>12.2</td>
<td>26.8</td>
<td>-3.8</td>
<td>-0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>Medit. California</td>
<td>±27.0</td>
<td>-45.4</td>
<td>-18.6</td>
<td>-3.6</td>
<td>9.8</td>
<td>23.0</td>
<td>-4.7</td>
<td>-1.0</td>
<td>1.3</td>
</tr>
<tr>
<td>NW Forested Mtns*</td>
<td>±45.6</td>
<td>-83.3</td>
<td>-35.0</td>
<td>-6.7</td>
<td>10.5</td>
<td>23.8</td>
<td>-13.0</td>
<td>-1.7</td>
<td>4.3</td>
</tr>
<tr>
<td>N. American Deserts*</td>
<td>±27.0</td>
<td>-33.2</td>
<td>-13.7</td>
<td>4.7</td>
<td>21.1</td>
<td>36.6</td>
<td>3.8</td>
<td>-0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>So. Semiarid Highlands</td>
<td>±30.0</td>
<td>-40.5</td>
<td>-7.2</td>
<td>4.9</td>
<td>19.0</td>
<td>45.2</td>
<td>5.8</td>
<td>-0.7</td>
<td>1.5</td>
</tr>
<tr>
<td>Temperate Sierras</td>
<td>±37.2</td>
<td>-64.6</td>
<td>-15.5</td>
<td>3.0</td>
<td>14.9</td>
<td>31.4</td>
<td>0.4</td>
<td>-1.4</td>
<td>2.2</td>
</tr>
<tr>
<td>Great Plains*</td>
<td>±18.2</td>
<td>-23.2</td>
<td>-9.1</td>
<td>3.9</td>
<td>14.8</td>
<td>23.6</td>
<td>2.5</td>
<td>-0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>E. Temperate Forests*</td>
<td>±14.3</td>
<td>-19.9</td>
<td>-10.4</td>
<td>-0.4</td>
<td>8.8</td>
<td>17.2</td>
<td>-0.5</td>
<td>-0.2</td>
<td>-0.1</td>
</tr>
<tr>
<td>Northern Forests</td>
<td>±18.9</td>
<td>-24.3</td>
<td>-13.0</td>
<td>-0.2</td>
<td>12.6</td>
<td>23.9</td>
<td>0.4</td>
<td>-0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Tropical Wet Forests</td>
<td>±24.6</td>
<td>-35.5</td>
<td>-16.7</td>
<td>0.5</td>
<td>15.2</td>
<td>27.4</td>
<td>-0.7</td>
<td>-0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

*Primary focus ecoregions
Figure 1: Distribution of the ten ecoregions covering the CONUS study domain.
Figure 2: Mean annual a) standard deviation (σ) and b) coefficient of variation (CV) calculated between CONUS-extent modeled estimates of hydrologic components precipitation (P), evapotranspiration (ET), runoff (R), snow water equivalent (SWE), soil moisture in equivalent water depth (SM(e)), and soil moisture in volumetric water content (SM(v)). Whiskers on each bar represent the standard deviation of annual values.
Figure 3: Boxplots of annual water year magnitudes of all study model estimates averaged over the CONUS extent for hydrologic components precipitation (P), evapotranspiration (ET), runoff (R), snow water equivalent (SWE), soil moisture in units of equivalent water depth (SM(e)), and soil moisture in units of volumetric water content (SM(v)). Dataset magnitudes are subdivided into two periods, 1985-1999 (left) and 2000-2014 (right). Flux component estimates (P, ET, R) are summed from monthly values to calculate annual water year rates. Storage component estimates (SWE, SM(e), SM(v)) are averaged from monthly values to calculate annual water year average storage values. Box lower limits, midlines, and upper limits represent the 25th, 50th (median), and 75th percentiles, respectively, of the associated data. Whiskers represent 1.5 times the interquartile range. Box color denotes dataset categories of hydrologic model, reanalysis, or remote sensing-derived. Asterisks are used to identify hydrologic models types: land surface models (** ** ** **), catchment models (** **), water balance models (**), and miscellaneous (*).
Figure 4: Forest plot of inter-annual (water year) component estimate Sen’s slope from 1982-2010 for components of precipitation (P), evapotranspiration (ET), snow water equivalent (SWE), soil moisture in units of equivalent water depth (SM(e)), and soil moisture in units of volumetric water content (SM(v)). Insignificant trends (p > 0.05) are grey. Significant trends are colored based on direction: negative trends are gold, positive trends are green. Datasets without complete data during the study period are represented with hollow points and temporal extent is noted under “Years in Period” column.
Figure 5: Boxplots of uncertainty, measured as annual coefficient of variation (CV), between component estimates at ten ecoregions. CV distributions are subdivided into “late” and “early” periods, 1985-1999 and 2000-2014. Box lower limits, midlines, and upper limits represent the 25th, 50th (median), and 75th percentiles, respectively, of the associated data. Whiskers represent 1.5 times the interquartile range. Box colors denote ecoregion and correspond to map colors in Figure 1. Components displayed are precipitation (P), evapotranspiration (ET), runoff (R), snow water equivalent (SWE), soil moisture in units of equivalent water depth (SM(e)), and soil moisture in units of volumetric water content (SM(v)). Results for SWE uncertainty are omitted for ecoregions with limited annual snowpack.
Figure 6: Bar plots summarizing the disagreement in Mann-Kendall trend (τ) direction using the unalikeability coefficient (u) after categorizing τ as significant negative trend, significant positive trend, or no significant trend for ten ecoregions from 1982-2010. Trends were assumed to be significant if \( p < 0.05 \). u can range from complete agreement (\( u = 0 \)) to complete disagreement (\( u = 1 \)) when sample size (n) is less than the number of categorical values (c). Because here n > c, maximum possible u is approximately 0.71 (dotted red line). Text to the right of each bar shows, in order, the u value and the number of datasets showing negative/insignificant/positive trend. Bars are color-matched to ecoregion colors from Figure 1 and are ordered from west (top) to east (bottom). Results are provided for water balance components of precipitation (P), evapotranspiration (ET), runoff (R), snow water equivalent (SWE), and soil moisture (SM) in units of either equivalent water depth or volumetric water content.
Figure 7: Spearman’s rho (ρ) correlation values of hydrologic model and reanalysis dataset component estimates against remote sensing-derived products for components of precipitation (P), evapotranspiration (ET), soil moisture in units of both equivalent water depth (SM(e)) and volumetric water content (SM(v)), and snow water equivalent (SWE). The title of each sub-plot (e.g. Figure 7a) provides the hydrologic component (e.g. P) and the remote sensing product against which the dataset was correlated (e.g. GPCP-v3). Statistically significant ρ values (p < 0.05) are shown as coloured circles. Insignificant ρ values (p > 0.05) are shown as black squares. Horizontal bars within each ecoregion denote mean ρ. Point and bar colours correspond to ecoregions, as shown in Figure 1. Each point or square corresponds to a single modeled dataset, with points sorted by descending ρ value. More detailed information can be found in Figure A2.6.
Figure 8: Histograms of ten ecoregion water budget relative imbalances calculated from combinations of all annual precipitation (15), evapotranspiration (15), and runoff (13) datasets with complete data from water years 2001-2010 (Table 1). Each region yielded 2,925 water budgets. Relative imbalances are calculated as a percentage of annual P. The 10th and 90th percentiles of each ecoregion distribution are provided as vertical dotted lines. Dashed vertical lines represent the relative imbalances of 10-year (2001-2010) water budgets from the ensemble mean of all modeled products.
Figure A2.1: Boxplots of annual standard deviation ($\sigma$) between component estimates at ten ecoregions. $\sigma$ distributions are subdivided into two periods, 1985-1999 and 2000-2014. Box lower limits, midlines, and upper limits represent the 25th, 50th (median), and 75th percentiles, respectively, of the associated data. Whiskers represent 1.5 times the interquartile range. Box colors denote ecoregions and correspond to map colors in Figure 1. Components displayed are precipitation (P), evapotranspiration (ET), runoff (R), snow water equivalent (SWE), soil moisture in units of equivalent water depth (SM(e)), and soil moisture in units of volumetric water content (SM(v)).
Figure A2.2: Boxplot summarizing annual peak snow water equivalent (SWE) depths by ecoregion derived from the distribution of modeled estimates used in this paper. Box lower limits, midlines, and upper limits represent the 25th, 50th (median), and 75th percentiles, respectively, of the associated data. Whiskers represent 1.5 times the interquartile range.
Figure A2.3: Heatmaps showing the relative timing of the beginning of snow water equivalent (SWE) accumulation (A2.3a) and ablation (A2.3b) periods for two ecoregions from 1985-2014. Relative timing is the difference between the month of the beginning of a model's accumulation or ablation period and the mean of all other model's timing. The beginning of accumulation and ablation is defined as when the rate of change of SWE between months is greater than 1 mm/month or less than -1 mm/month, respectively.
Figure A2.4: Heatmaps summarizing the annual relative timing values of SWE accumulation (accum.) and ablation from A2.3 for two ecoregions. Grid cells are colored by the percentage of years that have a positive relative timing (green) or by the percentage of years that have a negative relative timing (purple), whichever is greater.
### Comparison of Model Trends by Ecoregion

#### a) Precipitation

<table>
<thead>
<tr>
<th>τ ≥ 0.5</th>
<th>0 &lt; τ &lt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>Model</td>
</tr>
<tr>
<td>NW Forests</td>
<td>CA</td>
</tr>
<tr>
<td>M. Amer. Deserts</td>
<td>So., Semi-arid Highlands</td>
</tr>
</tbody>
</table>

#### b) Evapotranspiration

<table>
<thead>
<tr>
<th>τ ≥ 0.5</th>
<th>0 &lt; τ &lt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>Model</td>
</tr>
<tr>
<td>NW Forests</td>
<td>CA</td>
</tr>
<tr>
<td>M. Amer. Deserts</td>
<td>So., Semi-arid Highlands</td>
</tr>
</tbody>
</table>

#### c) Runoff

<table>
<thead>
<tr>
<th>τ ≥ 0.5</th>
<th>0 &lt; τ &lt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>Model</td>
</tr>
<tr>
<td>NW Forests</td>
<td>CA</td>
</tr>
<tr>
<td>M. Amer. Deserts</td>
<td>So., Semi-arid Highlands</td>
</tr>
</tbody>
</table>

#### d) Snow Water Equivalent

<table>
<thead>
<tr>
<th>τ ≥ 0.5</th>
<th>0 &lt; τ &lt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>Model</td>
</tr>
<tr>
<td>NW Forests</td>
<td>CA</td>
</tr>
<tr>
<td>M. Amer. Deserts</td>
<td>So., Semi-arid Highlands</td>
</tr>
</tbody>
</table>

#### e) Soil Moisture (equivalent water depth)

<table>
<thead>
<tr>
<th>τ ≥ 0.5</th>
<th>0 &lt; τ &lt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>Model</td>
</tr>
<tr>
<td>NW Forests</td>
<td>CA</td>
</tr>
<tr>
<td>M. Amer. Deserts</td>
<td>So., Semi-arid Highlands</td>
</tr>
</tbody>
</table>

#### f) Soil Moisture (volumetric content)

<table>
<thead>
<tr>
<th>τ ≥ 0.5</th>
<th>0 &lt; τ &lt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>Model</td>
</tr>
<tr>
<td>NW Forests</td>
<td>CA</td>
</tr>
<tr>
<td>M. Amer. Deserts</td>
<td>So., Semi-arid Highlands</td>
</tr>
</tbody>
</table>
Figure A2.5: Binned distributions of inter-annual model trends from 1982-2010 measured with Mann-Kendall’s tau (τ) for ten ecoregions, for water balance components of precipitation, evapotranspiration, runoff, snow water equivalent, and soil moisture. Model names are binned into very positive (τ ≥ 0.50, dark green), positive (0 < τ < 0.50, light green), negative (-0.50 < τ < 0, light purple), and very negative (τ ≤ -0.50, dark purple) significant trends, as well as an “insignificant” trend box (p > 0.05, grey). Trend significance is assumed when the p-value does not exceed 0.05 (p < 0.05). Model categories are denoted with a single letter identifying either a hydrologic (H), reanalysis (R), or remote sensing (S) model.
Figure A2.6: Distribution of hydrologic model and reanalysis dataset correlation against remote sensing products using Spearman’s rho ($\rho$). Values of $\rho$ are binned by $>0.90$, 0.50–0.90, $<0.50$, and statistically insignificant ($p > 0.05$). Rectangles are color-paired with their associated ecoregions (Fig. 1). Water budget components shown are precipitation (precip), evapotranspiration (ET), snow water equivalent (SWE), and soil moisture (SM).