AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR030966

Can We Use the Water Budget to Infer Upland Catchment Behavior? The Role of Data Set Error Estimation and Interbasin Groundwater Flow

Beatrice L. Gordon[1,2], **Wade T. Crow**[3], **Alexandra G. Konings**[4], **David N. Dralle**[5], and **Adrian A. Harpold**[1]

[1]Department of Natural Resources and Environmental Sciences, University of Nevada, Reno, Reno, NV, USA, [2]Graduate Program of Hydrologic Sciences, University of Nevada, Reno, Reno, NV, USA, [3]USDA-ARS Hydrology and Remote Sensing Laboratory, Beltsville, MD, USA, [4]Department of Earth System Science, Stanford University, Stanford, CA, USA, [5]Pacific Southwest Research Station, United States Forest Service, Davis, CA, USA

Key Points:

- Closed water budget assumptions about groundwater flux and measurement error are widely unsupported in upland catchments
- Triple Collocation (TC) indicates groundwater export in high arid catchments with deep substrates consistent with a framework by Fan (2019, https://doi.org/10.1002/wat2.1386)
- The Fan (2019, https://doi.org/10.1002/wat2.1386) framework, TC-based merging, and independent estimates of evapotranspiration advance upland water budget understanding

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

B. L. Gordon,
blgordon@alumni.stanford.edu

Citation:

Gordon, B. L., Crow, W. T., Konings, A. G., Dralle, D. N., & Harpold, A. A. (2022). Can we use the water budget to infer upland catchment behavior? The role of data set error estimation and interbasin groundwater flow. *Water Resources Research*, *58*, e2021WR030966. https://doi.org/10.1029/2021WR030966

Received 3 AUG 2021
Accepted 2 SEP 2022

Author Contributions:

Conceptualization: Wade T. Crow, David N. Dralle, Adrian A. Harpold
Formal analysis: Wade T. Crow, Adrian A. Harpold
Funding acquisition: Adrian A. Harpold
Investigation: Wade T. Crow, Alexandra G. Konings, Adrian A. Harpold

Abstract Water budgets are essential for characterizing water supplies from snow-dominated upland catchments where data are sparse, groundwater systems are complex, and measurements are prone to error (ε). One solution is imposing water budget closure (CWB) by ignoring difficult-to-measure variables, including inter-basin groundwater fluxes (G) and ε. However, conventional CWB-based analyses, which derive evapotranspiration (ET) from precipitation (P) and streamflow (Q) (e.g., the Budyko hypothesis), are limited in their ability to take advantage of recent advances in ET products, physically-based frameworks for improving inferences about G, or tools to statistically characterize ε (Triple Collocation [TC]); all of which offer promise for improved water supply predictions via open water budgets (OWB). We clarify the value of these advances in upland settings by comparing standard land surface model, Ensemble Mean, and TC-Merged P and ET products in 114 upland catchments. When compared against a long-term OWB, we find that the CWB assumptions are unsupported in 75%–100% of our 114 catchments, depending on the product. We then show how applying these CWB assumptions in snowy, steep catchments where ε is large can inflate inferences about streamflow response to climate change by up 9 times more than independent (OWB) estimates of ET using TC. Finally, we demonstrate how advances in OWB analysis reveal that high, arid settings with deep permeable substrate are groundwater exporters while most other basins are groundwater importers. Our results highlight the advantages of OWB analyses that harness new products, tools, and frameworks for characterizing inter-basin groundwater fluxes in critical upland settings.

Plain Language Summary Hydrologists use calculations about the inputs, outputs, and stores of water to anticipate changes in water supply from higher elevation, snow-dominated (upland) regions. To avoid challenges related to data limitations, it is common to make simplifying assumptions about negligible measurement error and groundwater storage and flow to close the annual water budget. However, the implications of these assumptions are rarely assessed. We investigate the support for these assumptions over a 15-year period and find that when invalid, they can over inflate the expected sensitivity of streamflow to precipitation falling as snow. We filter some error through optimized averaging (using a method called triple collocation) to illustrate how an open water budget approach that allows for measurement error and groundwater storage and flow can improve our understanding of water resources. This approach illustrates that higher, drier catchments with a deep subsurface provide a groundwater subsidy to lower catchments that is not apparent using other common techniques.

1. Introduction

Higher elevation (upland) catchments are critical for generating downstream water supplies (Barnett et al., 2008; Ehsani et al., 2017; Harpold et al., 2012; Li et al., 2017; Viviroli et al., 2007). As such, there is a pressing need to quantify how upland water supplies—including both surface and groundwater—will respond to changing climate (Gordon et al., 2022; Immerzeel et al., 2020; Mankin et al., 2015; Qin et al., 2020). Water budgets are foundational tools in this pursuit (Barnhart et al., 2016; Berghuijs et al., 2014; Ni et al., 2015). However, accurately closing the water budget in upland settings remains elusive due to complex hydrologic pathways and difficult-to-measure variables such as groundwater fluxes (G) and evapotranspiration (ET). Moreover, measurement error—

10.1029/2021WR030966

Methodology: Wade T. Crow, Adrian A. Harpold
Project Administration: Adrian A. Harpold
Resources: Wade T. Crow, Alexandra G. Konings, David N. Dralle, Adrian A. Harpold
Supervision: Wade T. Crow, Adrian A. Harpold
Validation: Wade T. Crow
Visualization: Wade T. Crow, Alexandra G. Konings, David N. Dralle, Adrian A. Harpold
Writing – original draft: Wade T. Crow, Alexandra G. Konings, David N. Dralle, Adrian A. Harpold
Writing – review & editing: Wade T. Crow, Alexandra G. Konings, David N. Dralle, Adrian A. Harpold

particularly biases (i.e., systematic error) in estimates of precipitation (P) associated with complex topography and snow plague water balance closure in upland catchments (Bales et al., 2006; Carroll et al., 2019; Henn et al., 2018). To circumvent these challenges, conventional approaches often assume that unknown or uncertain variables (including error) can be ignored by imposing a closed water budget (CWB, Fan, 2019; Kampf et al., 2020; Safeeq et al., 2021). To illustrate this point, we present a simple, but complete, catchment water budget following Fan (2019):

$$P - ET - Q = \frac{\Delta S}{\Delta t} + G + \varepsilon_P + \varepsilon_{ET} + \varepsilon_Q, \tag{1}$$

where P is precipitation, ET is evapotranspiration, Q is streamflow, $\frac{\Delta S}{\Delta t}$ is change in terrestrial water storage, G is interbasin-groundwater flux into or out of the catchment, and ε_P, ε_{ET}, and ε_Q are combined systematic and random error in the measurement of P, ET, and Q, respectively. Here, we define systematic errors as correlated to the true value of the variable whereas random errors are uncorrelated to the true value of the variable. Because they are difficult to separate, we refer to the combination of ε_p, ε_{ET}, and ε_Q as a single term (ε) and assume that ε is a combination of systematic and random error. When a CWB is imposed, right-hand side terms in Equation 1 including G, $\frac{\Delta S}{\Delta t}$, and ε are assumed to be zero, resulting in a simplified CWB form of the water budget in which $P–ET–Q = 0$. This simplification implies that ET can be calculated as $ET_{\mathrm{CWB}} = P–Q$.

Of the simple water budget-based tools that rely on CWB assumptions to derive ET_{CWB}, the Budyko hypothesis (Budyko, 1974) has emerged as a particularly useful and common framework for characterizing upland water resources (Barnhart et al., 2016; Berghuijs et al., 2014; Greve et al., 2020). The Budyko hypothesis posits that water budget partitioning (e.g., the ET fraction or ET/P and runoff ratio or Q/P) can be determined solely based upon the ratio of available energy (often expressed as annual potential evapotranspiration, E_o, in equivalent water depth) to available water (expressed as annual P, also in water depth) (Sposito, 2017). Relying on this simplification, prior research has attributed observations of systematic patterns (i.e., statistically detectable trends) in catchment plotting relative to Budyko-type curves to underlying physical processes or catchment properties (e.g., Li et al., 2013; Padrón et al., 2017; Potter et al., 2005), such as streamflow response to shifts in precipitation phase measured as the annual fraction of P falling as snow (fs) and intensity under climate change (Barnhart et al., 2016; Berghuijs et al., 2014; Ni et al., 2015).

To accurately characterize the water supplied from upland catchments, the above applications of the Budyko hypothesis require the imposition of (valid) CWB assumptions—specifically, that $G = 0$, $\varepsilon = 0 \rightarrow ET_{\mathrm{CWB}} = P–Q$. However, there is growing evidence that these assumptions may not hold in upland catchments (Kampf et al., 2020; Safeeq et al., 2021). First, G—specifically interbasin groundwater exportation—is typically ignored in CWBs, but can be a significant component of the upland water budget (Frisbee et al., 2016; Safeeq et al., 2021) with potential to bias ET_{CWB} either high or low (Fan, 2019). Prior research has sought to characterize how non-zero groundwater fluxes and stores can impact catchment plotting in the Budyko space using models and densely instrumented catchments (Condon & Maxwell, 2017; Istanbulluoglu et al., 2012; Wang et al., 2009). Quantifying interbasin groundwater import or export, however, is notoriously difficult in upland settings due to a lack of data and the complexity of hydrologic pathways (Carroll et al., 2019; Fan, 2019; Maxwell & Condon, 2016). In the absence of available data, Fan (2019) proposed a new physically-based framework comprised of several explanatory criteria to condition expectations about the role of G in water budgets including: (a) catchment size, (b) catchment position, (c) aridity, (d) depth of permeable regolith, and (e) geological permeability (see Section 3.2). For example, the framework argues that headwater catchments with deep permeable regolith are likely to export groundwater ($G > 0$), violating CWB assumptions. Second, ε—particularly systematic error in P in snowy, steep upland catchments—can also impact upland water budget closure, violating CWB assumptions and impacting ET_{CWB} as a result. Upland gauges are often too sparse to represent spatial variability in P (Jing et al., 2017) and are plagued by under-catch bias in snowy and windy conditions (Rasmussen et al., 2012). Moreover, orographic enhancement and complex terrain can hamper both satellite retrievals and high-resolution models in these settings (Dettinger, Cayan et al., 2014; Dettinger, Redmond, & Cayan, 2004; He et al., 2019; Henn et al., 2018; Wrzesien et al., 2019). Given the above, there is a pressing need to better understand how evidence for non-zero G and ε interacts with CWB assumptions to influence inferences about upland water supplies derived from widely used tools like the Budyko hypothesis (Andréassian & Perrin, 2012; Valéry et al., 2010).

Open water budget (OWB) approaches, which require an independent estimate of ET (henceforth, ET_{OWB}), offer a pathway to more rigorously investigate the role poor assumptions about G and ε may play in inferences about upland water supplies (Kampf et al., 2020). In upland settings, advances in land data assimilation systems (LDASs) (e.g., Kumar et al., 2019) and remote sensing (e.g., Anderson et al., 2011; Mu et al., 2011) have given rise to a suite of new ET_{OWB} datasets. However, each of these ET_{OWB} data sets are subject to considerable uncertainties (Polhamus et al., 2013), which makes it challenging for users to determine if the benefits of adopting ET_{OWB} outweigh the previously described disadvantages of propagating assumptions about G and ε into ET_{CWB}. If users adopt an OWB-based approach at all, they are often left to rely on best judgment in selecting from a range of ET_{OWB} data sets. In some cases, users might simply select an ET_{OWB} data set that uses other water budget variables like P and Q as inputs, hoping that this will reduce the total independent error sources (e.g., Barnhart et al., 2016). In other cases, users might elect to merge different estimates of ET_{OWB} together via an ensemble mean (e.g., Abolafia-Rosenzweig et al., 2021; Yilmaz et al., 2012). Doing so, however, implicitly assumes all estimates are equally error-prone. Information about the levels of errors in different ET_{OWB} estimates could facilitate a more accurate averaged best estimate. This can be achieved using triple collocation (TC)—a tool that objectively obtains random error estimates from three or more spatially and temporally collocated products to attain a single product with reduced random error (Gruber et al., 2017; Stoffelen, 1998; Yilmaz et al., 2012). Recent research has highlighted the value of TC for comparing different ET data sets (Khan et al., 2018). By improving estimates of P (Alemohammad et al., 2015), other research suggests that TC may also improve ET_{CWB} (Burnett et al., 2020) with additional benefits to water budget evaluations more generally. Despite the promise of TC, it is limited to providing statistical descriptions of random errors (see Section 3.1) and its limitations to describe or remove systematic errors impacting Equation 1 in upland settings are unexplored.

OWBs—aided by recent advances in approaches to condition expectations about G (e.g., Fan (2019)), tools to characterize ε (e.g., TC) and products to estimate ET_{OWB}— offer promise in evaluating the benefits and limitations of using CWB assumptions to make inferences about critical upland water supplies (Kampf et al., 2020; Safeeq et al., 2021). Because the value of these advances remains largely speculative, this study aims to clarify what—if any—benefit they provide to potential users in 114 snow-dominated upland catchments that import and export G and have large potential for ε (Condon et al., 2020; Fan, 2019; Henn et al., 2018; Rasmussen et al., 2012; Wrzesien et al., 2019). Using our study catchments, we first investigate of the validity of CWB assumptions through an evaluation of long term water budget closure and the Fan (2019) framework. Second, we interrogate the consequences of improper CWB assumptions when inferring the response of upland surface water supplies to climate change—specifically upland streamflow response to changes in P. Here, we focus our analysis of these benefits on the Budyko hypothesis because it is a widely used CWB-based tool that has been used to set expectations about upland water resources in response to climate change (Barnhart et al., 2016; Berghuijs et al., 2014; Greve et al., 2020). We follow Condon and Maxwell (2017) who more rigorously evaluate the effects of CWB assumptions in the Budyko space using ET_{CWB} and ET_{OWB}. Third, we examine whether these advances improve inferences about upland groundwater resources ignored in conventional applications of CWB-based tools like the Budyko hypothesis or otherwise masked by products with large ε. To do this we identify splits in our data using splits (e.g., thresholds) identified through conditional inference trees—a type of unbiased recursive partitioning. Using these splits, we establish statistically-based rules to categorically classify our watersheds based on agreement with the Fan (2019) criterion. We distill the motivation of this study into three central questions:

1. Does long term OWB closure—assisted by the Fan (2019) framework, TC, and ET_{OWB} products—validate conventional CWB assumptions (i.e., $G = 0$, $\varepsilon = 0 \rightarrow ET_{\mathrm{CWB}} = P–Q$) in upland catchments?
2. In upland catchments where CWB assumptions are invalid (e.g., $\varepsilon \neq 0$ due to P bias in steep, snowy catchments $\rightarrow ET_{\mathrm{CWB}} \neq P–Q$), how do the Fan (2019) framework, TC, and ET_{OWB} products differ in Budyko-based inferences about water supplies (i.e., streamflow efficiency in response to changes in precipitation phase, fs)?
3. When ε is characterized using TC, can the Fan (2019) framework and ET_{OWB} products improve insights about G export or import in upland settings?

2. Study Area and Data

2.1. Study Area

We focus on a large collection of upland catchments compiled using a subset of the Catchment Attributes and Meteorology for Large Sample studies (CAMELs) database (Newman et al., 2015). The CAMELS database is comprised of data for 671 catchments distributed throughout CONUS and includes forcing data for P, E_o, and Q. To select for upland catchments with high measurement uncertainty, we used the same subset of 268 catchments

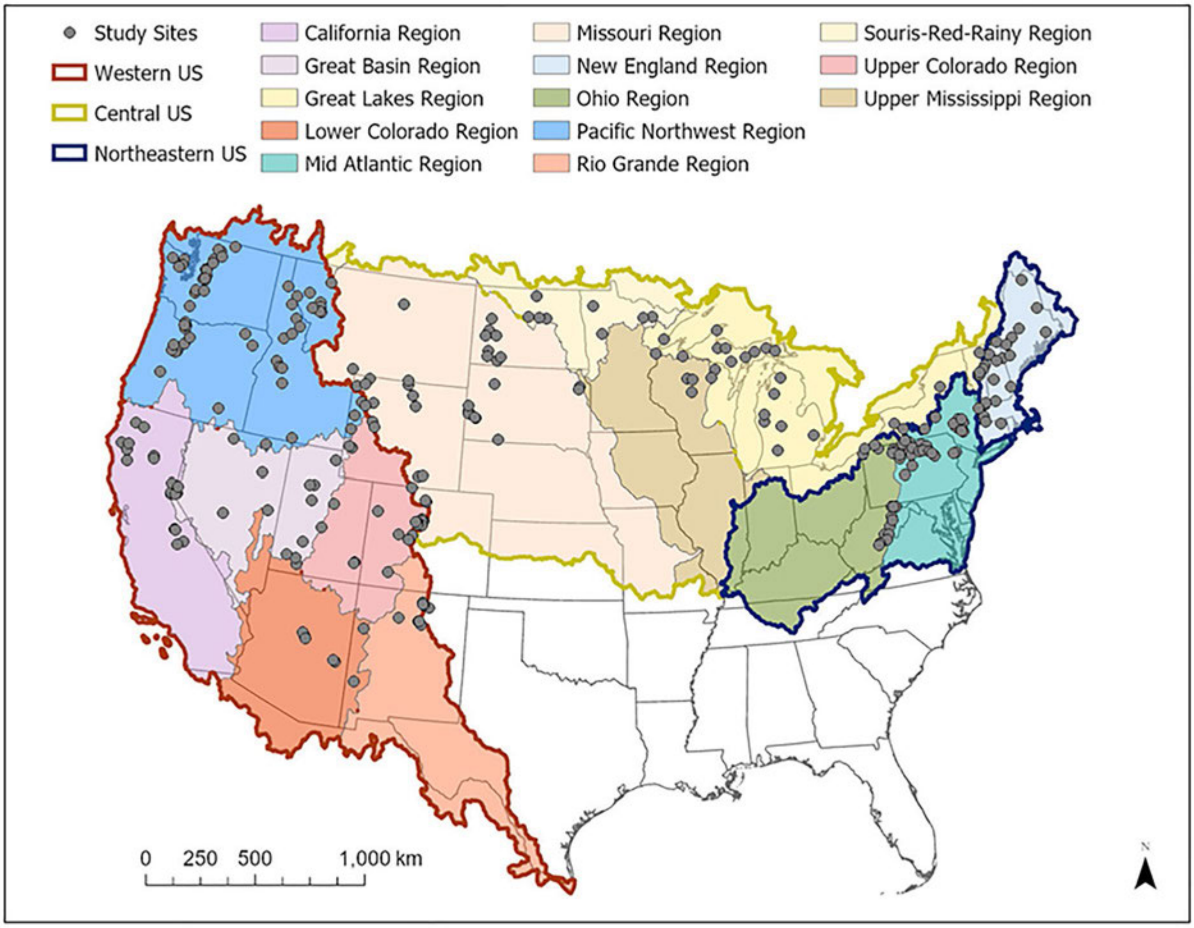

Figure 1. Map of candidate study sites broken out into the Western US (California, Great Basin, Pacific Northwest, Upper and Lower Colorado, and Rio Grande regions), Central US (Missouri, Upper Mississippi, Great Lakes, and Souris-Red-Rainey regions), and Northeastern US (Ohio, New England, and Mid-Atlantic regions). Here regions are abbreviated as follows: NE is New England, MA is Mid-Atlantic, OH is Ohio, GL is Great Lakes, UM is Upper Mississippi, MI is Missouri, SRR is Souris-Red-Rainey, UC is Upper Colorado, RG is Rio Grande, LC is Lower Colorado, GB is Great Basin, CA is California, and PNW is Pacific Northwest.

from the CAMELs database located throughout CONUS with $fs > 0.15$ used by Berghuijs et al. (2014). Catchments range from 6 to 2,679 km^2 in drainage area with a mean size of 387 km^2 (see Figure 1). Additional details on candidate catchments are provided in Table S1 in Supporting Information S1.

2.2. Data

In this study, we relied on multiple datasets (summarized in Table 1) to perform the analyses outlined in Section 3. For gridded time-series products, all data were obtained for the period from 1 October 2001 to 30 September 2016 for each catchment in Figure 1, which represents the maximum overlap between products allowable for the application of the TC approach outlined in Section 3.1. Please see Text S1 and S2 in Supporting Information S1 for more detail.

To assist users in evaluating the advantages and disadvantages associated with different estimates of ET_{OWB} and to test the value of TC-based merging in upland settings, we compared three different data products selected to represent the most common choices available to users in upland settings: (a) an NLDAS product (Section 2.2.1); (b) an Ensemble Mean product (Section 2.2.2); and (c) a TC-Merged product (described in Sections 2.2.3 and 3.1 below).

2.2.1. NLDAS Product (Example of a Single Model)

The NLDAS product consists of NLDAS P forcing data, which is derived from a temporal disaggregation of gauge-only Climate Prediction Center (CPC) data with the PRISM topographical adjustment, CPC hourly

Table 1
Summary of Input Data Products Used for This Study Including Five P Products and Four ET Products

Product	Spatial resolution	Spatial extent	Temporal resolution	Temporal extent	Source	Included in ensemble mean?	Reference
Water budget and Budyko variables							
Precipitation							
ERA5	31 km	Global	Hourly	1979-Present	Reanalysis	No	(Hersbach et al., 2020)
PERSIANN-CDR	27.75 km	Quasi-Global	Daily	1983-Present	Satellite	No	(Ashouri et al., 2015; Sorooshian et al., 2014)
NLDAS-2	13.875 km	North America	Hourly	1979-Present	Gauge Based	Yes	(Xia et al., 2012) Included in CAMELs database
Daymet	1 km	North America	Daily	1980 -Present	Gauge Based	Yes	(Thornton et al., 2016) Included in CAMELs database
PRISM	4 km	CONUS	Daily	1981-Present	Gauge Based	Yes	(Daly et al., 1997, 2008)
Evapotranspiration							
ALEXI	10 km	CONUS	Daily	2001-Present	Thermal: TSEB	No	(Anderson et al., 2011)
SSEBop	1 km	CONUS	Daily	2000-Present	Thermal: Penman-Monteith	Yes	(Senay et al., 2013)
NCA-LDAS	12 km	CONUS	Daily	1979–2016	Land Surface Model: Penman-Monteith	Yes	(Kumar et al., 2019)
MODIS16	0.5 km	Global	8-day	2001-Present	Satellite (Near Infrared): Penman-Monteith	Yes	(Mu et al., 2013)
Potential evapotranspiration[a]							
NLDAS-2	13.875 km	North America	Hourly	1979-Present	Penman Based	N/A	(Xia et al., 2012) Included in CAMELs database
Streamflow[a]							
USGS	catchment	North America	Daily	1980–2014	Gauge Based	N/A	(Addor et al., 2017; Newman et al., 2015) Included in CAMELs database
Catchment and subbasin attributes							
Snow Fraction							
–	catchment	North America	Average Annual	1980–2014 (CAMELS); variable	–	N/A	(Addor et al., 2017; Newman et al., 2015) Included in CAMELs database
Depth to bedrock							
Pelletier	catchment	Global	–	–	Model-Based	N/A	(Pelletier et al., 2016 Included in CAMELs database
Maximum sub-basin elevation							
SRTM	3 arc s	Global	–	–	Model-Based	N/A	(Yang et al., 2011
Mean catchment aridity							
–	catchment	CONUS	Average Annual	1980–2014 (CAMELs); variable	Model-Based	N/A	(Addor et al., 2017; Newman et al., 2015) Included in CAMELs database
Geologic permeability							
GLHYMPS	~100 km^2 (average polygon size)	Global	–	–	Database synthesis	N/A	Gleeson et al., 2014 Included in CAMELs database

Note. Details about specific data products are summarized in Table 1 below. We note whether the data were included in the CAMELs database in the reference column. Data that were not available in the CAMELs database were independently estimated from the sources listed in Section 7. We provide more detail on each product in Text S2 of Supporting Information S1.

[a]Supplemented with additional data as described in Text S2 in Supporting Information S1.

CONUS gauge data, hourly Doppler radar precipitation data, half-hourly CPC data, 3-hourly North American Regional Reanalysis data (Xia et al., 2012). The NLDAS product also includes NCA-LDAS *ET*, which is generated by running the National Climate Assessment-Land Data Assimilation System (LDAS) with NLDAS forcings, including *P* (Kumar et al., 2019). The NLDAS product was selected to investigate the relative advantage of selecting a single data set for *P* and *ET* (e.g., Barnhart et al., 2016). NCA-LDAS was specifically selected based on the widespread use of LDASs for agricultural and water resources management applications and based on NCA-LDAS' explicit focus on the terrestrial water cycle. Unlike the other products, errors in NLDAS products are not mutually independent and are constrained to balance (over sufficiently long time periods). That is, unlike the two products below, for the NLDAS product errors in *P* may be compensated by errors in *Q* and/or *ET*.

2.2.2. Ensemble Mean Product (Example of Merging of Multiple Data Sets)

The Ensemble Mean product in this study was comprised of an ensemble mean of NLDAS, PRISM, and Daymet *P* data sets and an ensemble mean of MOD16, NCA-LDAS, and SSEBop *ET* data sets, which are noted in Table 1. The Ensemble Mean product was selected to investigate the common approach of merging multiple data sets. We elected to combine the *P* and *ET* products into the Ensemble Mean above based on their public availability, ease of access, and widespread use in studies of upland/mountain environments (Addor et al., 2017; Hahm, Dralle et al., 2019; Hahm, Rempe et al., 2019; Newman et al., 2015; Velpuri et al., 2013).

2.2.3. TC-Merged Product (Example of Optimized Merging of Multiple Data Sets)

The TC-Merged product in this study was comprised of all *P* and *ET* data sets in Table 1 and was constructed following the methodology outlined in Section 3.1 below. The TC-Merged product was selected to investigate the merging of multiple datasets using an objective statistical representation of their random errors. Our constructed *P* triplets followed prior work (Massari et al., 2017) and were comprised of the following: (a) a reanalysis product; (b) a satellite-based product; and (c) a gauge-based interpolation or gauge-based interpolation/Land Surface Model (LSM) hybrid. *ET* triplets were constructed using: (a) a thermal product, (b) a LSM; and, (c) a near-infrared product.

3. Methods

To answer our three research questions, we first constructed new estimates of *ET* and *P* using TC-based merging (Section 3.1.), which required a robust characterization of cross-correlated estimation errors (Section 3.1.1). Once we characterized cross-correlated errors, we used TC outputs as the basis for optimized merging (Section 3.1.2) to obtain TC-Merged *P* & *ET* with minimized random error (Section 2.2.3). We then evaluated the performance of TC-Merged *P* & *ET* against NLDAS (Section 2.2.1) and Ensemble Mean (Section 2.2.2) *P* & *ET*.

We used TC-Merged, NLDAS, and Ensemble Mean *P* & *ET* to assess the validity of CWB assumptions, which are necessary for conventional application of the Budyko hypothesis. We did this by comparing "inferred groundwater behavior" (e.g., $G+\varepsilon$) obtained using each of our three products in a long term OWB. We evaluated the long term OWB by rearranging Equation 1 as:

$$P_{\text{sum}} - ET_{\text{sum}} - Q_{\text{sum}} = G_{\text{sum}} + \varepsilon_{\text{sum}} \tag{2}$$

where the subscript "sum" indicates the 15-year sum of each variable. We neglected the contribution of $\frac{\Delta s}{\Delta t}$ because existing measurements of it were too coarse for our application (Tapley et al., 2004) and it was challenging to reliably separate G from $\frac{\Delta s}{\Delta t}$ (Enzminger et al., 2019). Consistent with reported common assumptions for USGS streamflow data from Hamilton and Moore (2012), we assumed Q_{sum} was accurate to within 5% at a 95% confidence interval. We normalized the resulting inferred groundwater behavior (i.e., $G+\varepsilon$) by *P* and sorted catchments into the three different categories described in Table 2: groundwater neutral, groundwater importer, or groundwater exporter.

Because each category in Table 2 can be influenced by true G and/or ε, we further tested the physical support for inferred groundwater behavior using five criteria proposed by Fan (2019) (described further in Section 3.2). We applied a type of unbiased recursive partitioning to establish splits in the five criteria related to differences in inferred groundwater behavior and then used the raw number of supporting criteria as the basis for classifying catchments with physically supported groundwater import or export (Section 3.2.1). Next, we used these results to examine the impacts of propagating poor CWB assumptions about G and ε into ET_{CWB} via the Budyko hypothesis (Section 3.3).

10.1029/2021WR030966

Table 2
Summary of Water Budget Closure Categories Adopted for This Study Based on Fan (2019) and Their Implications for G and ε

Category	Mathematical description per Equation 2	Implication for $G_{\mathrm{sum}} + \varepsilon_{\mathrm{sum}}$	Implication for CWB
Groundwater neutral	$P_{\mathrm{sum}} - ET_{\mathrm{sum}} - Q_{\mathrm{sum}} = 0$	$G_{\mathrm{sum}} + \varepsilon_{\mathrm{sum}} = 0$	$ET_{\mathrm{CWB}} = ET_{\mathrm{OWB}}$
Groundwater exporter	$P_{\mathrm{sum}} - ET_{\mathrm{sum}} - Q_{\mathrm{sum}} > 0$	$G_{\mathrm{sum}} + \varepsilon_{\mathrm{sum}} > 0$	$ET_{\mathrm{CWB}} > ET_{\mathrm{OWB}}$
Groundwater importer	$P_{\mathrm{sum}} - ET_{\mathrm{sum}} - Q_{\mathrm{sum}} < 0$	$G_{\mathrm{sum}} + \varepsilon_{\mathrm{sum}} < 0$	$ET_{\mathrm{CWB}} < ET_{\mathrm{OWB}}$

Note. P_{sum} is the sum of P over the 15-year period of record, ET_{sum} is the sum of ET over the 15-year period of record, and so on.

3.1. Characterization of Random ε: TC Analysis

TC analysis requires the construction of a TC triplet with three distinct measurement systems (e.g., X, Y, and Z) of the same environmental variable (McColl et al., 2014; Stoffelen, 1998). Henceforth, we use TC[X-Y-Z] to refer to a generic TC triplet of a single environmental variable constructed from measurement systems X, Y, and Z. In order to obtain valid outputs for TC[X-Y-Z], TC requires that a number of assumptions are met (Gruber et al., 2017). The first of these assumptions (assumption 1) is that each measurement system or product (e.g., X, Y, and Z) has a linear relationship to the "true" variable (C_T) which can be modeled for a generic variable C as follows:

$$C_i = \alpha_i C_T + \epsilon_i \tag{3}$$

Here, I is the individual measurement system or product (e.g., X, Y, or Z), α_i is a measure of the relation between C_i and C_T, and ϵ_i are the respective random zero-mean errors associated with each measurement system or product. P errors are typically modeled as multiplicative in short-term and/or fine-scale applications (Alemohammad et al., 2015); however, recent work by Massari et al. (2017) and Dong et al. (2019) suggests that the assumption of a multiplicative error model for TC application at the daily timescale is not necessary. As such, we assumed that underlying error model for both 8-day P and ET was linear and no logarithmic transformation was applied.

The second assumption (assumption 2) relates to signal and error stationarity, where stationarity is satisfied if the statistical properties of a geophysical process do not change over time. To satisfy assumption 2, we tested the performance of raw time-series data against time-series anomalies obtained by removing both the long-term and seasonal mean (e.g., DJF, MAM, JJA, and SON, Text S1 in Supporting Information S1). Because the TC results for all P and ET triplets were equivalent using raw and anomaly time-series data (not shown), we used raw time-series data in this analysis to avoid any effects introduced from seasonality determination on TC results.

The final two assumptions (assumptions 3 and 4) are that the signal and the error in the geophysical measurements are independent (error orthogonality) and that the errors in the selected geophysical measurements are independent (zero error cross-correlation) (Gruber et al., 2016). Additionally, obtaining valid extended triple collocation (ETC) results requires that each triplet have at least 50 or more points and that all correlation outputs are positive for each of the input timeseries data sets (Chen et al., 2018). If assumptions 1–4 are satisfied, then the error variances (σ^2) can be determined using six unique elements from sample covariance matrix (e.g., A_{XY}) between three measurement systems using TC following Equation (4–6):

$$\sigma_{\mathrm{x}}^2 = A_{\mathrm{XX}} - \frac{A_{\mathrm{XY}} A_{\mathrm{XZ}}}{A_{\mathrm{YZ}}} \tag{4}$$

$$\sigma_{\mathrm{y}}^2 = A_{\mathrm{YY}} - \frac{A_{\mathrm{XY}} A_{\mathrm{YZ}}}{A_{\mathrm{XZ}}} \tag{5}$$

$$\sigma_{\mathrm{z}}^2 = A_{\mathrm{ZZ}} - \frac{A_{\mathrm{XZ}} A_{\mathrm{YZ}}}{A_{\mathrm{XY}}} \tag{6}$$

The additional contribution of ETC following McColl et al. (2014), is the estimation of Pearson's correlation coefficient (R) between C_i and C_T using Equation 7, which estimates R between C_X and C_T as an example, and where σ_{XY} is the covariance between C_X and C_Y:

$$R_{\mathrm{X}} = \sqrt{\frac{\sigma_{\mathrm{XY}} \sigma_{\mathrm{YZ}}}{\sigma_{\mathrm{x}}^2 \sigma_{\mathrm{XY}}}} \tag{7}$$

3.1.1. Estimation of Uncertainties and 95% Confidence Interval Calculations

Because of the 15-year period of record available across all data, considerable estimation errors arising from differences in our sample and the true variable were expected in our results. Furthermore, because two of the data products used were incorporated in all possible triplets, estimation errors were also assumed to be highly correlated across products (Chen et al., 2018). This correlation affects the R_X calculated by ETC, despite R_X representing a value that is only influenced by product X (C_X) and the true signal C_X. Comparison of R_X values across different triplets can thus be used to detect the influence of estimation errors on the TC results (Crow et al., 2017; Yilmaz & Crow, 2014). For example, with minimal bias from cross-correlated estimation errors, the R_X obtained from TC[X-Y-Z] would be expected to be the same (or very similar) to R_X obtained from TC[X-Y-W]. We would therefore anticipate small pairwise differences in calculated correlation values (ΔR). Thus, it is important to quantify when the difference in pairwise ΔR values is significant to appropriately interpret ETC results.

To detect cross-correlated estimation errors, we obtained uncertainty intervals for ΔR values sampled across all catchments using a 1000-member boot-strapping approach. Pairwise ΔR were assessed in this manner for all common products (e.g., the R_X obtained from TC[X-Y-Z]–R_X obtained from TC[X-Y-W]). Each boot-strapped sample in this approach was constructed using the exact same set of days. We then constructed 95% confidence intervals from the boot-strapped sampling distribution, which we defined as the range between the 2.5th and 97.5th percentile of boot-strapped ΔR values. We assumed that the results for any catchment that fell outside the 95% confidence interval in any pairwise ΔR comparison were unacceptably impacted by cross-correlated estimation errors and removed those catchments from further analysis.

3.1.2. TC-Based Merging

We then used a composite of the TC-based merging methodologies put forward by Yilmaz et al. (2012) and Gruber et al. (2017) to obtain an estimate of the true underlying values for both P and ET based on the ETC results. This method was also applied by Burnett et al. (2020). Using this methodology, TC-based error variance estimates (as opposed to ETC-based R estimates) can be used to obtain a single more accurate data set for a given geophysical variable (denoted C_M, the TC-based merged estimate of C_T):

$$C_M = w_x C_X + w_y C_Y + w_z C_Z, \tag{8}$$

where $w$is a weight obtained from the TC-based error variances obtained from Equations 4–6 obtained for each measurement system or product using Equations 9–11:

$$w_x = \frac{\sigma_y^2 \sigma_z^2}{\sigma_x^2 \sigma_y^2 + \sigma_x^2 \sigma_z^2 + \sigma_y^2 \sigma_z^2} \tag{9}$$

$$w_y = \frac{\sigma_x^2 \sigma_z^2}{\sigma_x^2 \sigma_y^2 + \sigma_x^2 \sigma_z^2 + \sigma_y^2 \sigma_z^2} \tag{10}$$

$$w_z = \frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 \sigma_y^2 + \sigma_x^2 \sigma_z^2 + \sigma_y^2 \sigma_z^2} \tag{11}$$

3.2. Characterization of G: Fan (2019) Analysis

Because both G and/or ε can contribute to inferred groundwater behavior (Table 2) based on the evaluation of Equation 2, we investigated the physical support for apparent groundwater neutrality, export, or import in each catchment using five criteria proposed by Fan (2019):

- **Criterion 1 (Catchment Scale)**. Small catchment size increases the likelihood of groundwater importer or exporter behavior. If the catchment is small compared to its permeable regolith, the catchment has a greater likelihood of being a groundwater exporter. We assessed this factor in our study catchments using catchment size sourced from the CAMELS database and depth to bedrock estimates from Pelletier et al. (2016).
- **Criterion 2 (Catchment Position)**. Catchments situated at the high end of a regional elevation gradient are more likely to be groundwater exporters (Buss et al., 2013) while catchments situated at the low end of the

gradient are more likely to be importers (Genereux et al., 2013). We assessed this factor in our study catchments using the ratio of mean catchment elevation relative to the maximum elevation of the sub-basin (HUC8) containing each study catchment (e.g., higher ratio corresponds to higher catchment position and vice versa). Mean catchment elevation was assessed using data from the CAMELs database and maximum sub-basin elevation was assessed using data from Yang et al. (2011).

- **Criterion 3 (Climate).** Headwater catchments under a dry climate with seasonal aridity or interannual droughts are more likely to be groundwater exporters. Fan (2019) argue that an arid climate leads deep local water tables below stream beds in the headwaters. Water from precipitation and losing streams in arid headwater catchments enters regional groundwater flow systems and resurfaces in lower basins to feed gaining systems (Käser & Hunkeler, 2016). They further hypothesize that the amount of groundwater export will increase if the climate is wetter in the headwaters and drier in the lower basin and decrease in the reverse case. We assessed this factor in our study catchments using aridity data from the CAMELs database (Addor et al., 2017).
- **Criterion 4 (Substrate Properties).** Catchments underlain by thick regolith, fractured rock, or sediments in the headwaters facilitate are likely to be groundwater exporters (Buss et al., 2013). This is particularly true in tectonically active terrain under a humid climate. We assessed this factor using depth to bedrock from Pelletier et al. (2016).
- **Criterion 5 (Geologic Structure).** Catchments situated atop high-permeability, dipping sedimentary beds extending beyond the catchment, or shared sedimentary beds—particularly carbonate rocks—are more likely to be groundwater importers or exporters. We assessed the geologic permeability (expressed as mu(K) where K is hydraulic conductivity using data from the GLobal HYdrogeology MaPS (GLHYMPS) (Gleeson et al., 2014).

For each data product, we first assessed the statistical relationships between Fan (2019) criteria and inferred groundwater behavior based on Equation 2 using Kruskal-Wallis and pairwise Wilcoxon rank sum tests. We binned catchments into roughly equal groups unless there were objective breakpoints (e.g., climate and substrate properties). We then used a Kruskal-Wallis test to determine if any statistically significant differences were observed between groups and OWB closure assuming an α of 0.05. In the case that statistically significant differences were observed using the Kruskal-Wallis test, we then used a pairwise Wilcoxon rank sum test to further characterize the statistical significance of differences between groups.

To relate the Fan (2019) criteria to our data, we established splits (e.g., thresholds) in each Fan (2019) variable used for categorical classification of our catchments. Splits were determined using conditional inference trees, which are a type of unbiased recursive partitioning (Hothorn, Hornik, & Zeileis, 2012; Hothorn, Hornik, Van De Wiel, & Zeileis, 2012). Following Hothorn et al. (2015), the algorithm we used first tested independence between input variables (e.g., catchment scale, position, and so on) and the response (i.e., OWB closure). The algorithm stopped if all variables are found to be independent of the response; otherwise, the variable with the strongest association —as measured by a p-value —was selected and a binary split was implemented. This process was then repeated recursively to obtain splits reported in Table 3. For more details on this method we refer to Hothorn et al. (2015) and for more details on its implementation in this study we refer to Figures S1–S6 in Supporting Information S1. We report the qualitative Fan (2019) criterion, the split identified by unbiased recursive partitioning that was assumed to represent any qualitative threshold (e.g., catchment was positioned on the high end of a regional gradient) reported by Fan (2019), and the resulting rules for defining agreement between apparent groundwater behavior and the Fan (2019) criteria in Table 3.

3.2.1. Classification of G or ε Dominance Analysis

Fan (2019) argued that multiple criteria provide the strongest evidence for groundwater leakage; in other words, satisfying one of the criteria outlined in Section 3.2.2. Does not unambiguously support groundwater export or import. For example, they find that catchments are more likely to be leaky if positioned at the high end of a steep regional gradient, underlain by deep substrates, and in a drier climate. As such, we do not focus on whether a catchment satisfies a single criterion for our classification. We grouped catchments into four categories based on the agreement between inferred groundwater behavior based on the evaluation of Equation 2 and agreement with the Fan (2019) criteria as defined in Table 3.

- **Self-Contained.** We defined catchments as self-contained if long term $G_{\mathrm{sum}} + \varepsilon_{\mathrm{sum}}$ was 0% ± Q uncertainty bounds. These catchments were not assessed for agreement with Fan (2019) criteria.

Table 3

Summary of the Classification Rules for Fan (2019) Factors Developed via Conditional Inference Trees as Described by Hothorn et al. (2015) and Reported in Figures S1–S6 in Supporting Information S1

Criterion	Split based on unbiased recursive partitioning	Agreement rule with Fan (2019)
Catchment scale	No statistically significant relationship (Figure S1 in Supporting Information S1)	None defined
Position	Statistically significant split in the relationship between OWB closure and position (m/m) > 0.56 (Figure S2 in Supporting Information S1)	Apparent groundwater export if OWB + position >0.56 Apparent groundwater import if OWB + position <0.56
Climate	Statistically significant split in the relationship between OWB closure and aridity (mm/mm) > 0.42 (Figure S3 in Supporting Information S1)	Apparent groundwater export if OWB + aridity> 0.42
Substrate properties	Statistically significant split in the relationship between OWB closure and soil depth (*m*) > 2.09 (Figure S4 in Supporting Information S1)	Apparent groundwater export if OWB + soil depth > 2.09
Geologic structure	Statistically significant split in the relationship between OWB closure and permeability (mu(K)) > −12.6 (Figure S5 in Supporting Information S1)	Fan (2019) does not indicate clear expectation for *G*, so none defined

Note. We define two rules for position as Fan (2019) clearly indicate that lower lying catchments are more likely to be importers and higher up catchments are more likely to be exporters. In the absence of agreement with any rules listed, catchments retained their apparent groundwater behavior (e.g., importer, exporter, or neutral), but were classified with lower confidence per Section 3.2.1.

- **Dominant ε (No Fan (2019) criterion met).** We assumed that εin the underlying water budget variables caused the lack of agreement between any explanatory physical criteria and inferred groundwater behavior. We had the lowest degree of confidence in the physical realism of our upland water budgets in these catchments.
- ***G* and ε (One Fan (2019) criterion met).** We assumed that both ε and G influenced the agreement between only one explanatory physical criterion and inferred groundwater behavior. We did not attempt to further disentangle the influence of ε and G due to data limitations and had a lower degree of confidence in the physical realism of our upland water budgets in these catchments.
- **Dominant *G* (Two or more Fan (2019) criteria met).** We assumed that agreement between multiple explanatory physical criteria and inferred groundwater behavior indicated true signal associated with G. We had the highest degree of confidence in the physical realism of our upland water budgets in these catchments.

3.3. Impacts of Unsupported CWB Assumptions in the Budyko Space Analysis

The impacts of broken CWB assumptions on water budget-based inferences are underexplored (Andréassian & Perrin, 2012; Koppa & Gebremichael, 2017). In order to investigate potential impacts, we first conducted a direct comparison of ET_{OWB} and ET_{CWB}using each of our three data products. We then assessed how broken CWB assumptions influenced deviation from a Budyko modeled ET (i.e., ET_{Budyko}) fraction by contrasting our observed ET_{CWB}fraction against our observed ET_{OWB} fraction. Finally, we evaluated how these differences impacted inferences about Budyko streamflow anomaly (Q_{anom}) to *fs*—a metric commonly used as a proxy for climate change in upland catchments (Gordon et al., 2022).

3.3.1. Budyko *ET* Fraction Analysis

We compared long term estimates of ET_{CWB} or ET_{OWB} across all study sites using our three different products. Using potential evapotranspiration data from NLDAS (Table 1, Text S2 in Supporting Information S1) and our three different realizations of P, we then evaluated how our different realizations of ET fraction interacted with G and ε in the Budyko space. To do this, we used the conventional form of the Budyko hypothesis presented in Equation 12 to obtain a modeled estimate of $ET_{\mathrm{Budyko}}$by multiplying the left-hand side by P:

$$\frac{\overline{ET_{\mathrm{Budyko}}}}{\overline{P}} = \sqrt{\frac{\overline{E_o}}{\overline{P}}\tanh\left(\frac{\overline{P}}{\overline{E_o}}\right)\left(1-\exp\left(-\frac{\overline{E_o}}{\overline{P}}\right)\right)}, \tag{12}$$

where $\overline{ET}$, $\overline{P}$, and $\overline{E_o}$ are mean annual ET, P, and E_o over the 15-year period of record and $\overline{ET_{\mathrm{Budyko}}}$ is the mean annual modeled Budyko ET.

3.3.2. Budyko Q_{anom} Fraction Analysis

In upland settings, analyses of Budyko streamflow anomalies (Q_{anom}) have indicated that greater-than-expected streamflow efficiency (or runoff ratio—Q/P) is correlated to higher *fs*, suggesting that streamflow will decline

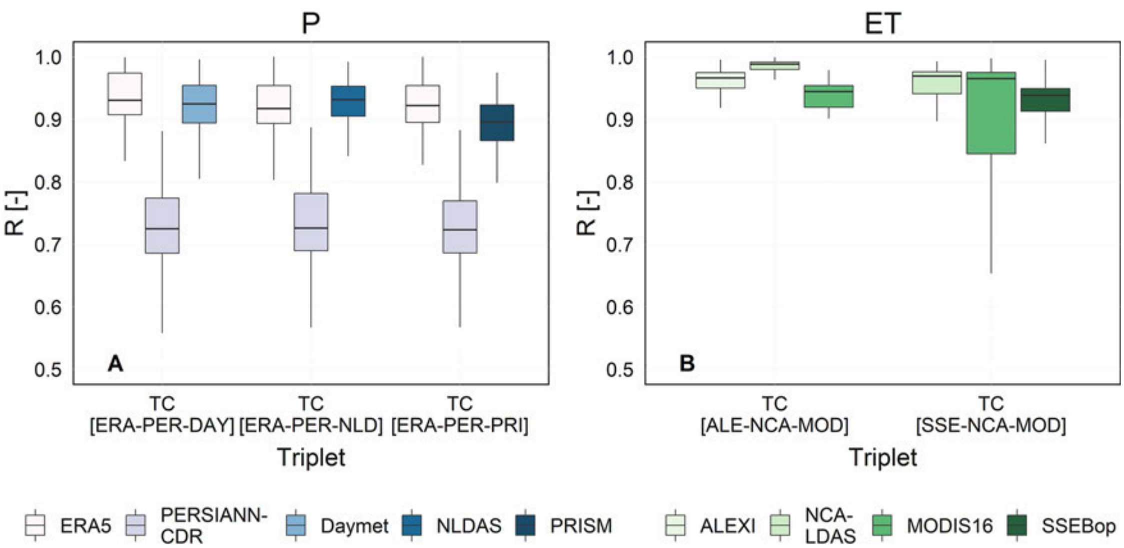

Figure 2. (a) Extended triple collocation (ETC) R results for P products in all valid catchments ($n = 170$); and (b) ETC R results for ET products in all valid catchments ($n = 143$). Boxplots describe variability across catchments. Here R is an estimation of the correlation between the product and the "true" underlying P or ET value as described in Section 3.1. Boxplots show the minimum, 25th percentile, median, 75th percentile, and maximum values. Catchments where any pairwise ΔR values (Section 3.1.1) fell outside the 95% confidence intervals (indicated by black dashed lines in Figure S7 and S8 of Supporting Information S1, respectively) were excluded (see Tables S2 and S3 in Supporting Information S1).

with an increase in winter rain under climate change (Berghuijs et al., 2014). However, to the best of our knowledge, these analyses have not considered whether findings are sensitive to assumptions about G and ε imbedded in ET_{CWB} despite evidence for large systematic ε due to P under-catch in snow-dominated catchments (Lundquist et al., 2021; Wrzesien et al., 2019). Here, we test how G and ε influence the relationship between Budyko streamflow anomaly and fs using the equation below:

$$\frac{\overline{Q_{\mathrm{anom}}}}{\overline{P}} = \left(1 - \frac{\overline{ET}}{\overline{P}}\right) - \left(1 - \frac{\overline{ET_{\mathrm{Budyko}}}}{\overline{P}}\right), \tag{13}$$

where $\overline{ET}$ is mean annual ET calculated via ET_{CWB} or ET_{OWB} over the 15-year period of record and $\overline{ET_{\mathrm{Budyko}}}$ is obtained from Equation 12. We then examined the correlation between different realizations of streamflow anomaly and mean annual fs.

4. Results

4.1. Characterization of Random ε: TC-Merging Results

A limitation of our TC-based analysis is the need to exclude catchments with cross-correlated estimation error (Section 3.1). Catchments where any pairwise ΔR values (Section 3.1.1) fell outside the 95% confidence intervals (indicated by black dashed lines in Figure S7 and S8 of Supporting Information S1, respectively) were excluded (see Table S2 and S3 in Supporting Information S1). This exclusion led to a reduction in the number of study catchments (from 268 to 114). In the valid 114 catchments, pairwise ΔR values were small and close to zero (see red lines in S7 and S8). The similarity in mean R values obtained for different products (e.g., ERA5 P in Figure 2a shaded in light purple) suggests that any bias from cross-correlated sampling errors in the constructed triplets had a small impact on ETC results. Based on this logic, ETC was deemed successful in 114 of the 268 candidate catchments. There were 143 valid catchments for ET and 170 catchments for P, with 114 catchments that had both. For successful catchments, individual performance—as measured by R—in P products (Figure 2a) was more varied than for ET (Figure 2b). The ERA5 most strongly correlated to the unknown "true" value of P in 85 of 170 catchments (~50%), NLDAS in 58 of 170 catchments (~34%), Daymet in 20 of 170 catchments (~12%), and PRISM in 7 of 170 catchments (~4%). In 124 of 143 catchments (~87%), NCA-LDAS exhibited the strongest correlation to the unknown "true" value of ET. In the remaining 19 catchments, ALEXI had the strongest correlation to the true ET in 8 catchments (~6%), MOD16 in 6 catchments (~4%), and SSEBop in 5 catchments (~3%).

Table 4
Summary of Inferred Groundwater Behavior (e.g., $G_{sum} + \varepsilon_{sum}$) Based on the Evaluation of Equation 2 With NLDAS, Ensemble Mean, and TC-Merged Products

		NLDAS		Ensemble mean		TC-merged	
Category	Mathematical description	#	% of total	#	% of total	#	% of total
Groundwater neutral	$G_{sum} + \varepsilon_{sum} = P_{sum} - Q_{sum} - ET_{sum} = 0$	25	21.9%	0	0%	29	25.4%
Groundwater exporter	$G_{sum} + \varepsilon_{sum} > P_{sum} - Q_{sum} - ET_{sum} > 0$	7	6.1%	0	0%	33	28.9%
Groundwater importer	$G_{sum} + \varepsilon_{sum} < P_{sum} - Q_{sum} - ET_{sum} < 0$	82	71.9%	114	100%	52	45.6%

4.2. Characterization of G: Fan (2019) Results

We assessed Equatio 2 using TC-Merged P & ET, NLDAS P & ET, and Ensemble Mean P & ET to obtain an estimate of inferred groundwater behavior (i.e., $G + \varepsilon$). Median inferred groundwater behavior was −1.8% of P using TC-Merged P & ET, −9.7% of P using NLDAS P & ET, and −19.3% of P using Ensemble Mean P & ET; all of which are consistent with findings by Safeeq et al. (2021) in a smaller number of densely instrumented catchments. Table 4 reports the number of catchments classified as groundwater exporters (i.e., $P_{sum}–Q_{sum}–ET_{sum} > 0$), groundwater neutral (i.e., $P_{sum}–Q_{sum}–ET_{sum} \cong 0$), or groundwater importers (i.e., $P_{sum}–Q_{sum}–ET_{\mathbf{sum}} < 0$) in accordance with the definitions outlined in Table 2. A greater number of catchments were characterized as groundwater exporters or groundwater neutral in the Northeastern and Central US (e.g., green and tan boxes in the top half of Figure 3) using TC-Merged P & ET. Conversely, all products indicated widespread regional groundwater importation in the Western US (e.g., red and blue colored boxes in the bottom half of Figure 3) and particularly the Pacific Northwest (labeled PNW in Figure 3) although this could be error related per our results in Figure 5.

We found that OWB closure using TC-Merged P & ET was the most consistent with physical reasoning based on Fan (2019). For example, when TC-Merged P & ET were used we observed that groundwater exporters were positioned higher up in the containing sub-basin (Figure 4f), were variably more arid (Figure 4i), and had deeper permeable regolith or fractured rock (Figure 4l), consistent with Fan (2019). TC-Merged P & ET indicated statistically significant differences between inferred groundwater behavior and catchment position (Table 5), with groundwater exportation (i.e., positive $G + \varepsilon$) observed for higher catchment positions (Figure 4f) as postulated by Fan (2019). Although NLDAS P & ET indicated a similar rightward shift in the boxplots as drainage position increased (Figure 4d), it supported widespread groundwater importation (i.e., negative $G + \varepsilon$) counter to our expectation. TC-Merged P & ET also supported a statistically significant relationship between higher groundwater exportation (i.e., positive $G + \varepsilon$) and aridity (Figure 4i and Table 5). Conversely, NLDAS and Ensemble Mean P & ET indicated a statistically significant relationship between groundwater importation (i.e., negative $G + \varepsilon$) and aridity (Figures 4g and 4f, Table 5). Consistent with the expectation that deep regolith or fractured rock increased groundwater export (Fan, 2019), TC-Merged P & ET supported statistically significant relationships between greater depth to bedrock and groundwater export (i.e., positive $G + \varepsilon$) (Figure 4l). Neither NLDAS nor Ensemble Mean P & ET (Figure 4j–4l) indicated any statistically significant relationships between depth to bedrock and inferred groundwater behavior (Table 5).

4.2.1. Classification of G or ε Dominance Results

We separated catchments into the four groups reported in Table 6 using the classification rules established via our unbiased recursive partitioning analysis (Table 3). Per Figure 5, we observed distinct regional patterning in our catchment classification. For example, of the 37 catchments located in the Western US (Table S2 in Supporting Information S1) that were not apparently self-contained, there was strong physical support for inferred groundwater behavior in only one catchment (~3% of 37) with weak support for 10 catchments (~27% of 37) and inconclusive support in 23 catchments (~62% of 37) using TC-Merged P & ET with NLDAS and Ensemble Mean P & ET yielding similar results. Some of these catchments in the Western US with the greatest implied ε were in steeper terrain and generally had a larger fs (consistent with gauge under-catch; see Figure S9 in Supporting Information S1), which would be expected to lead to the systematic under-prediction of P (Henn et al., 2018; Rasmussen et al., 2012; Wrzesien et al., 2019). Overall, TC-Merged P & ET yielded the highest number of catchments with strong physical support as assessed against the Fan (2019) framework, particularly in the Central and Eastern US (Figure 5, Table 6).

10.1029/2021WR030966

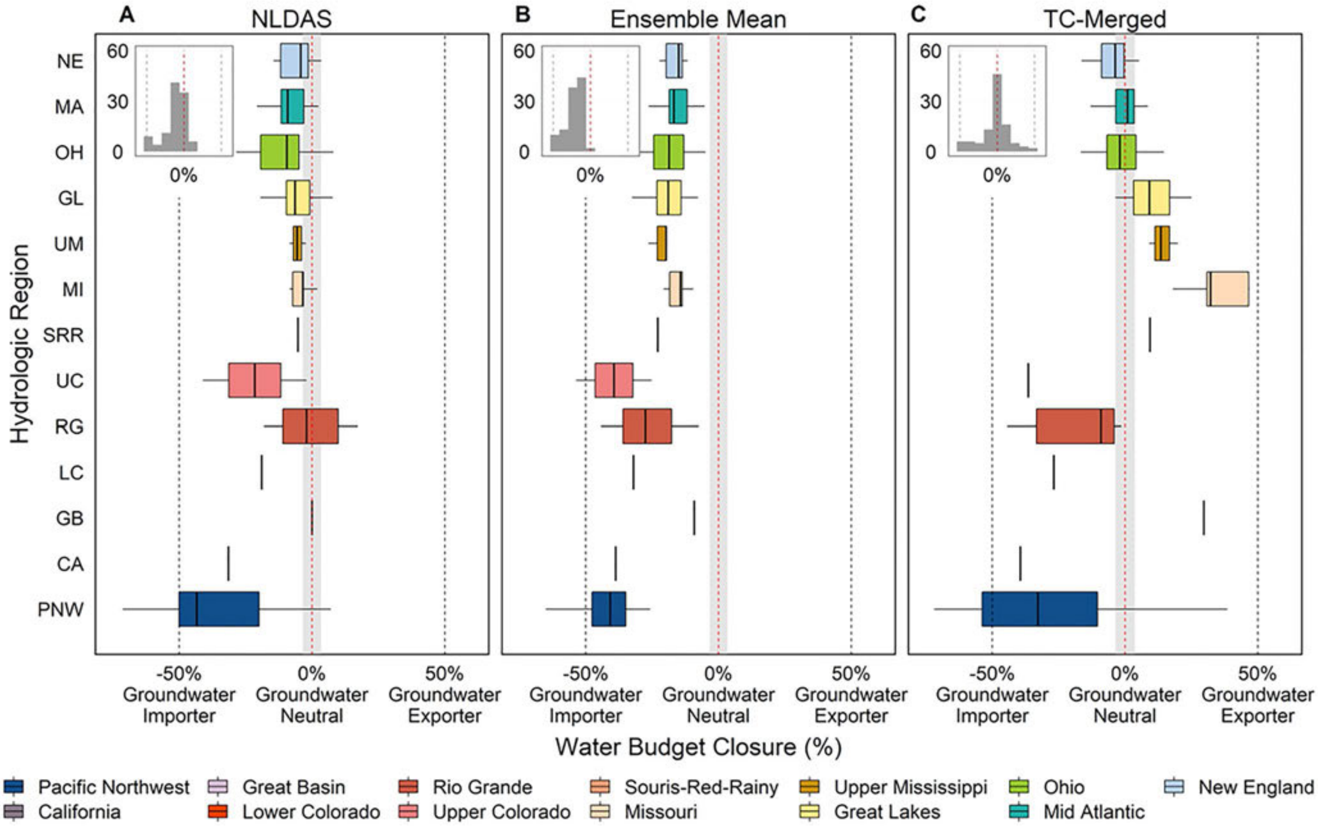

Figure 3. Regional variability in the evaluation of Equation 2 where: (a) groundwater importers are defined based on long-term water budget closure support for negative G_{sum} ($P_{sum} - Q_{sum} - ET_{sum} < 0$), groundwater neutrality is defined based on long-term water budget closure support for G_{sum} equal to 0 ($P_{sum} - Q_{sum} - ET_{sum} \cong 0$), and groundwater exporters are defined based long-term water budget support for positive G_{sum} ($P_{sum} - Q_{sum} - ET_{sum} > 0$) are estimated using the NLDAS P & ET; (b) same as above but with Ensemble Mean P & ET; (c) same as above but with TC-Merged P & ET. We plot the water budget closure as a percent of each respective P_{sum} to facilitate more intuitive and contextual interpretation of the results. Gray bounds indicate the potential uncertainty introduced by Q_{sum}. We refer the reader to Figure 1 for abbreviations. Inset histograms represent the distribution of open water budgets closure across catchments using each of the three different products.

4.3. Impacts of Unsupported CWB Assumptions

4.3.1. Budyko *ET* Fraction Results

Our results show that choices about *ET,* including whether it is evaluated independently (e.g., ET_{OWB}) or derived (e.g., ET_{CWB}) by ignoring G and ε, can lead to substantially different hydrologic inferences in upland catchments (Figures 6a–6c). Across all products, we observed limited agreement between ET_{CWB} and ET_{OWB} with inset histograms reinforcing widespread potential for ET_{CWB} to underestimate ET_{OWB} (warmer colors in Figures 6a–6c). This underestimation of ET_{OWB} was particularly pronounced in places (e.g., steep, snowy catchments in the western US, Figure S9 in Supporting Information S1) where the estimation of P is more variable (larger circles in Figures 6a–6c). We suggest that this is consistent with large unconsidered ε (Figure 5) arising from systematic under-prediction of P (Henn et al., 2018; Rasmussen et al., 2012; Wrzesien et al., 2019). TC-Merged P & ET were less obviously biased toward underestimation than either NLDAS or Ensemble Mean P & ET, particularly in the central and eastern US. Here, the TC-Merged P & ET indicated potential for ET_{CWB} to overestimate ET_{OWB}, which is consistent with unconsidered G exportation and our classification results (Figure 5).

In the Budyko space, we found that the largest differences in plotted *ET* fraction were driven by the propagation of CWB assumptions about G and ε into ET_{CWB} with smaller differences based on the selection of an individual product (Figures 7a, 7c and 7e vs. Figures 7b, 7d and 7f). When ET_{CWB} was used, we observed that *ET* fraction was substantially lower than expected based on Equation 12 in catchments impacted by ε (triangles and inverted triangles in 7A, C, and E) when compared to smaller deviations in catchments with physically realistic G (squares in Figure 7) or that were self-contained (circles in Figure 7) consistent with Jones et al. (2012). Voepel et al. (2011) suggested that steeper catchments may be associated with lower *ET* fraction; however, we found that

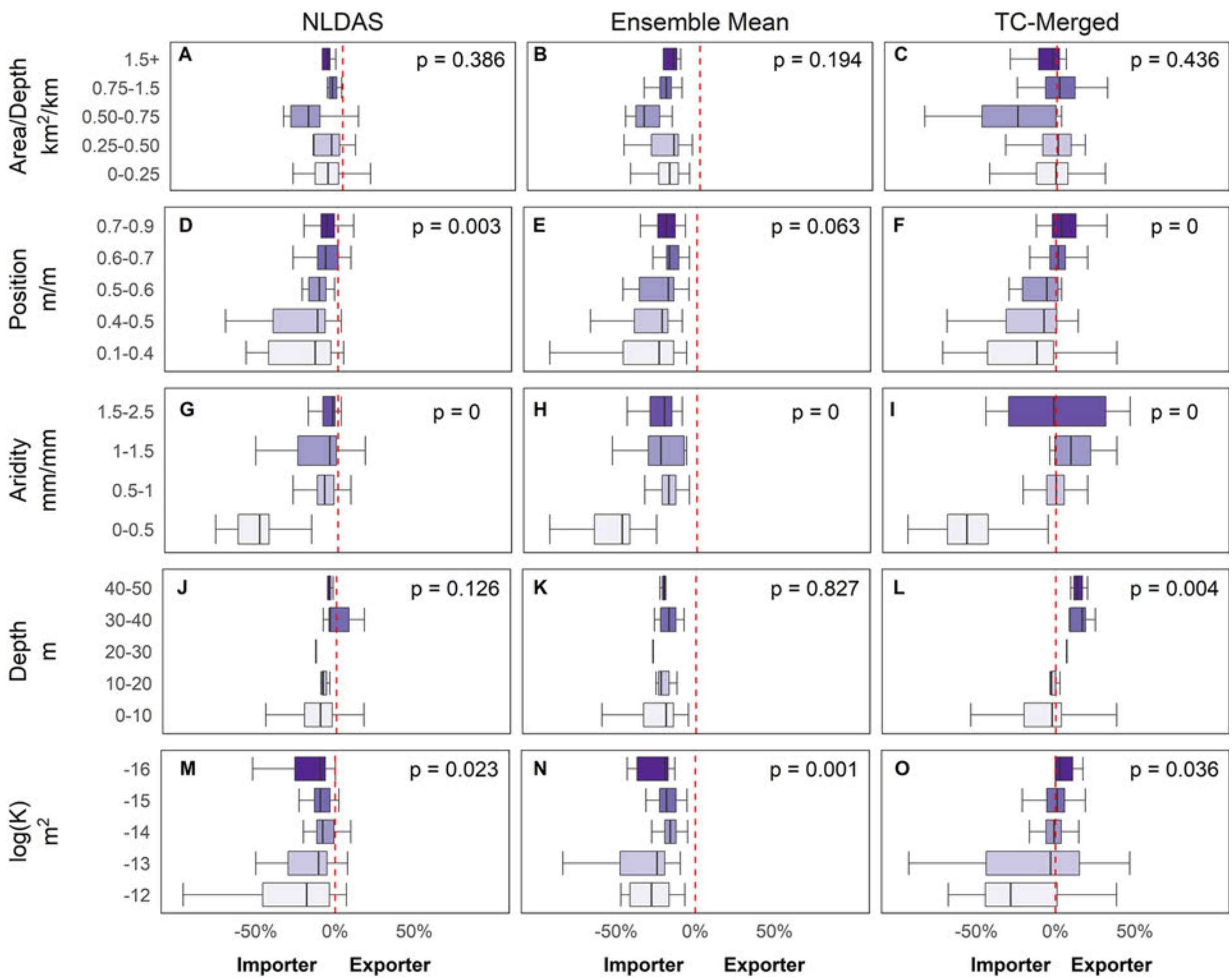

Figure 4. Boxplots showing the 25th percentile, 50th percentile, and 75th percentile and the standard error bars for inferred groundwater behavior based on Equation 2 versus each of the five Fan (2019) criteria proposed in Section 3.2.1 to determine whether a catchment is an importer, neutral, or exporter. The criteria are plotted as follows: (a–c) Catchment size relative to depth of permeable regolith (Criteria 1); (d–f) Catchment position relative to a regional gradient measured as the ratio of mean catchment elevation to maximum sub-basin elevation; –GI) Catchment climate as measured using the aridity index; j–l) Catchment depth of permeable substrate or fractured rock as measured by depth to bedrock; and m–o) Catchment geological permeability as measured by the log of hydraulic conductivity (k). Inferred groundwater behavior is approximated for plots a, g, d, j, and m using NLDAS P & ET, for plots b, e, h, k, and n using Ensemble Mean P & ET, and plots c, f, i, l, and o using TC-Merged P & ET. We report the Kruskal-Wallis p-value assuming a significance level of $\alpha = 0.05$ for each plot. If the Krushkal-Wallis p-value was statistically significant, we reported the results of a Pairwise Wilcoxon Rank Sum Test (Table 5).

steepness may increase ε bias in observed ET fraction. That is, catchments impacted by ε also tended to be steeper, snowier, and in the Western US (see Figure S9 in Supporting Information S1), where other research suggests that there is large potential for ε —particularly systematic P under-prediction (Henn et al., 2018; Rasmussen et al., 2012; Wrzesien et al., 2019). When ET_{OWB} was used, G and ε had less obvious influence over ET fraction (Figures 7b, 7d and 7f), underscoring how these assumptions are propagated into the Budyko space via ET_{CWB} when not properly considered.

4.3.2. Budyko Q_{anom} Results

Per Equation 13, lower-than-expected ET fraction drives higher-than-expected streamflow anomaly in Budyko-based assessments. Although previous research has attributed higher-than-expected streamflow anomaly to larger fs in upland catchments (Barnhart et al., 2016; Berghuijs et al., 2014; Ni et al., 2015), we found that the strength of this relationship is also systematically influenced by ε in particular when ET_{CWB} is used. For example, when TC-Merged P was used to obtain ET_{CWB}, we found that catchments impacted by ε (dashed lines in Figure 8e, Table 7) were 5.5 times more sensitive to fs than catchments with physically realistic G (dotted lines in Figures 8e and Table 7) and 3 times more sensitive than self-contained catchments (solid lines in Figure 8e, Table 7). This patterned response

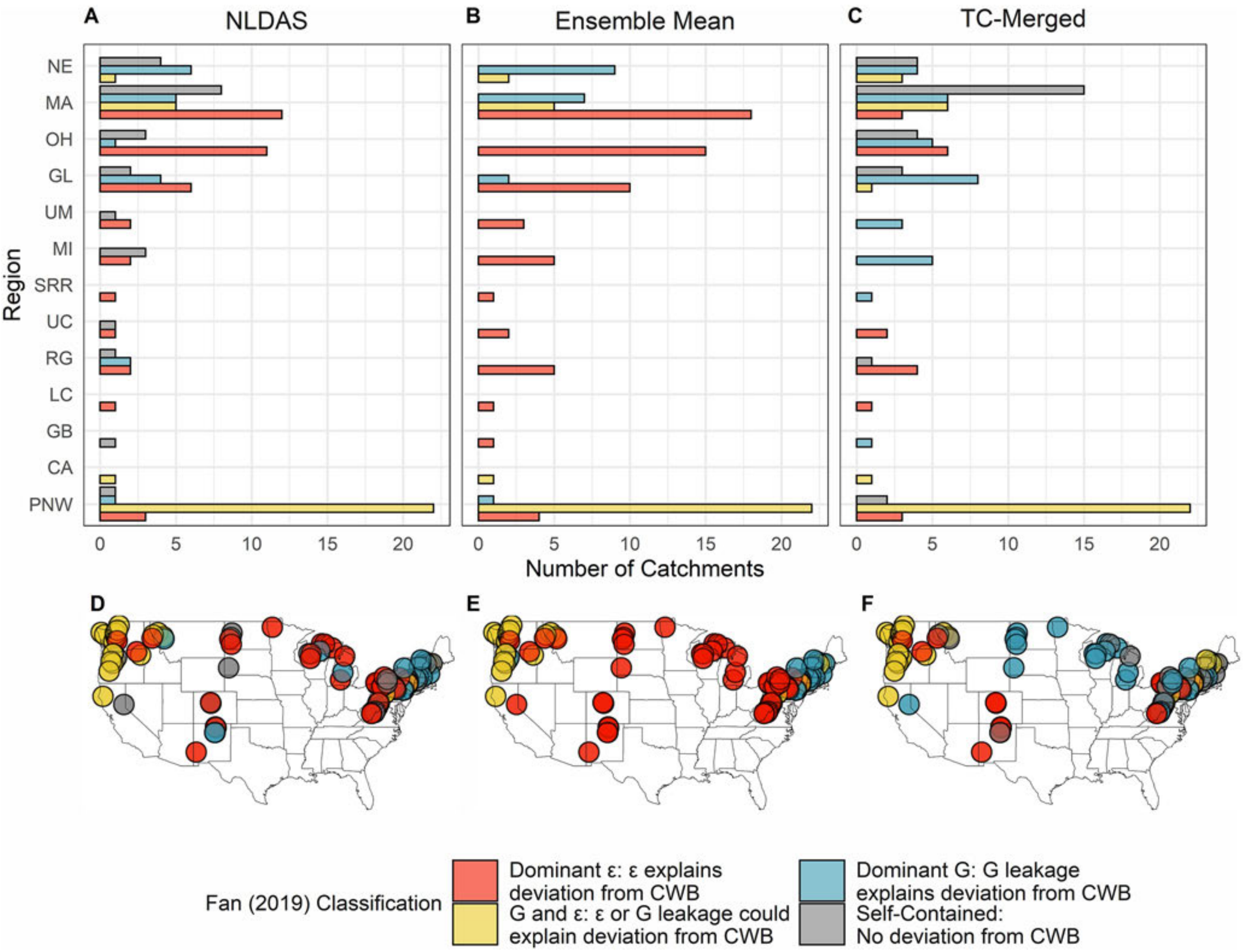

Figure 5. Variability in physical support for inferred groundwater behavior based on the Fan (2019) criterion presented in Section 3.2.1 using (a) NLDAS *P* & *ET*; (b) Ensemble Mean *P* & *ET*; and (c) TC-Merged *P* & *ET*. Maps displaying catchment classification using (d) NLDAS *P* & *ET*; (e) Ensemble Mean *P* & *ET*; and (c) TC-Merged *P* & *ET*. Abbreviations for the regions correspond to Figure 1.

was similar using NLDAS *P* (Figure 8c), but more muted using Ensemble Mean *P* (Figure 8a). The contrast with ET_{OWB} across all products further highlighted the effect of poor CWB assumptions about ε versus *G* on the relationship between streamflow anomaly and *fs* (Figures 8b, 8d and 8f, Table 7). Notably, when TC-Merged *P* & *ET* were used in catchments impacted by *G* (i.e., with strong physical support for underlying water budget), streamflow sensitivity to *fs* was consistent between ET_{CWB} (slope = 0.13, Table 7) and ET_{OWB} (slope = 0.11, Table 7) and lower than previous results (e.g., slope = 0.37 from Berghuijs et al. (2014) and 0.32 from Barnhart et al. (2016)). Both TC-Merged *P* & *ET* and NLDAS *P* & *ET* also yielded consistent slopes in self-contained catchments per Table 7.

5. Discussion

5.1. Does Long Term OWB Closure—Assisted by the Fan (2019) Framework, TC, and ET_{OWB} Products—Validate Conventional CWB Assumptions in Upland Catchments?

Neglecting the contributions of *G* and ε to upland water budgets by imposing closure conveniently reduces data requirements to *P* and *Q*. However, our results show that these assumptions are unsupported in a range of upland settings (see Figures 3 and 5) with significant implications for CWB-based tools, including conventional forms of the Budyko hypothesis. We show that when a CWB is inappropriately applied, the true *G* and ε magnitudes are deposited into a calculated *ET* (ET_{CWB}) (Figure 6), leading to systematic biases in catchment plotting in Budyko space (Figures 7 and 8) as discussed in Section 5.2. We contrast these results with an evaluation of Equation (2)

Table 5
Table of the Statistically Significant Results for Pairwise Wilcoxon Rank Sum Tests for Each of the Five Fan (2019) Factors Proposed in Section 3.2.1

Factor	Pairwise comparison	Group 1		Group 2		NLDAS	Ensemble mean	TC-merged
		Name	Sample size	Name	Sample size	*p* Value	*p* Value	*p* Value
Position (m/m)	0.1–0.4 versus 0.6–0.7	0.1–0.4	20	0.6–0.7	21	0.008	0.048	0.005
	0.1–0.4 versus 0.7–0.9	0.1–0.4	20	0.7–0.9	38	0.008	*NS*	0.000
	0.4–0.5 versus 0.6–0.7	0.4–0.5	19	0.6–0.7	21	0.012	0.008	0.017
	0.4–0.5 versus 0.7–0.9	0.4–0.5	19	0.7–0.9	38	0.006	0.044	0.001
	0.5–0.6 versus 0.6–0.7	0.5–0.6	16	0.6–0.7	21	0.047	*NS*	0.025
	0.5–0.6 versus 0.7–0.9	0.5–0.6	16	0.7–0.9	38	0.020	*NS*	0.002
Aridity (mm/mm)	0–0.5 versus 0.5–1	0–0.5	17	0.5–1	79	0.000	0.000	0.000
	0–0.5 versus 1–1.5	0–0.5	17	1–1.5	7	0.002	0.002	0.002
	0–0.5 versus 1.5–2.5	0–0.5	17	1.5–2.5	11	0.000	0.000	0.000
Depth (m)	0–10 versus 30–40*	0–10	102	30–40	5	*NS*	*NS*	0.003
	0–10 versus 40–50*	0–10	102	40–50	3	*NS*	*NS*	0.023
	10–20 versus 30–40	10–20	3	30–40	5	*NS*	*NS*	0.036
log(K) (m^2)	−12 versus −14	−12	21	−14	29	0.007	0.01	0.019
	−12 versus −15	−12	21	−15	31	0.013	0.005	0.003
	−12 versus −16	−12	21	−16	8	*NS*	*NS*	0.032
	−13 versus −14	−13	25	−14	29	0.024	0.001	*NS*
	−13 versus −15	−13	25	−15	31	*NS*	0.002	*NS*

Note. We assumed a significance level of $\alpha = 0.05$ and report significant p-values for pairwise comparisons within the groups above using a Pairwise Wilcoxon Rank Sum Test. We did not calculate pairwise statistics if the Kruskal-Wallis p-value was not significant.

using an OWB-based approach that leverages a physically-based framework for characterizing G proposed by Fan (2019), TC-based merging, and several independent estimates of ET_{OWB}. The combined result of these advances indicated widespread non-zero $G + \varepsilon$ in a range of upland settings: for example, inferred groundwater behavior (i.e., $G + \varepsilon$) as determined by the evaluation of Equation 2 using TC-Merged P & ET suggested that 85 of 114 upland catchments were not self-contained. These findings provide broad and quantitative evidence in support of recent calls to revisit common assumptions about G and ε used to close the water budget and derive ET_{CWB} in upland settings (Fan, 2019; Kampf et al., 2020; Safeeq et al., 2021).

At the same time, our results point to the very real challenge of disentangling signal associated with G from ε (Figures 3 and 5, Tables 4–6) in upland settings that lack comprehensive groundwater datasets (Fan, 2019) and are more susceptible to the systematic under-prediction of P (Lundquist et al., 2019; Rasmussen et al., 2012; Wrzesien et al., 2019). This is particularly true for steeper, snowier catchments in the western US due to orographic enhancement and complex terrain (Dettinger, Cayan et al., 2014; Dettinger, Redmond, & Cayan, 2004; He et al., 2019; Henn et al., 2018; Wrzesien et al., 2019), sparse precipitation measurements (Jing et al., 2017),

Table 6
Summary of Inferred Groundwater Behavior (e.g., $G_{sum} + \varepsilon_{Ssum}$) Based on the Evaluation of Equation 2 With NLDAS, Ensemble Mean, and TC-Merged Products

Classification	Agreement with Table 3 rules	Assumptions	NLDAS		Ensemble mean		TC-merged	
			#	% of total	#	% of total	#	% of total
Self-Contained	N/A	$G = 0$, $\varepsilon = 0$	25	21.9%	0	0.0%	29	25.4%
Dominant ε	No rules satisfied	Very weak physical support for inferred groundwater behavior	41	36.0%	65	57.0%	19	16.7%
G and ε	One rule satisfied	Inconclusive physical support for inferred groundwater behavior	29	25.4%	30	26.3%	33	28.9%
Dominant G	Two or more rules satisfied	Strong physical support for inferred groundwater behavior	19	16.7%	19	16.7%	33	28.9%

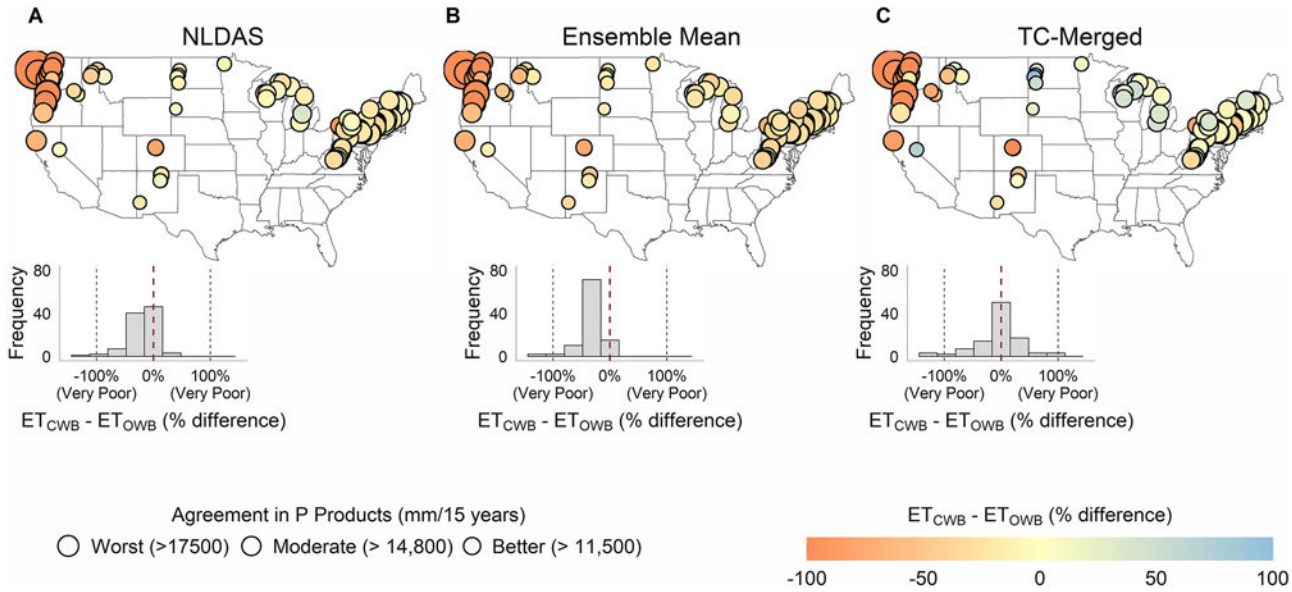

Figure 6. Differences between long-term ET_{CWB} and ET_{OWB} across all catchments (n = 114) using: (a) NLDAS P & ET and USGS Q; (b) Ensemble Mean P & ET and USGS Q; and (c) TC-Merged P & ET and USGS Q. Coloring is based on observed on the percent difference between ET_{CWB} and ET_{OWB} with values less than −100% and more than 100% constrained to those bounds for plotting. Sizing is based on the maximum disagreement between long-term estimates of P. We present scatterplots of ET_{CWB} versus ET_{OWB} colored by aridity in Figure S10 of Supporting Information S1.

and P under-catch (Rasmussen et al., 2012). Consistent with this body of research, we found that the evaluation of Equatiob 2 in these catchments was disproportionately impacted by ε (Figures 3 and 5, S9 in Supporting Information S1), limiting confidence in the resulting water budgets regardless of the products, approaches, and tools used. Although this challenge is a significant one, as bias correction methodologies improve estimations of P (Beck et al., 2020) the benefits of using TC-Merged estimates of ET_{OWB} to evaluate Equation 2 may enable more robust water budget-based supply predictions in these particular settings.

5.2. In Upland Catchments Where CWB Assumptions Are Invalid, How Do the Fan (2019) Framework, TC, and ET_{OWB} Products Differ in Budyko-Based Inferences About Water Supplies?

Recent research has highlighted the need to better connect systematic shifts in catchment location in the Budyko space to physical explanation (Berghuijs et al., 2020). Previous work (Hahm, Dralle et al., 2019; Hahm, Rempe et al., 2019; Istanbulluoglu et al., 2012; Jones et al., 2012) has suggested that groundwater losses can lead to systematic deviations in catchment plotting below the Budyko curve due to lower ET fractions and illustrated impacts to streamflow (Han et al., 2021), with a diminished focus on ε. Meanwhile, other work has examined the influence of ε without consideration of G (Andréassian & Perrin, 2012; Koppa & Gebremichael, 2017; Valéry et al., 2010). Our results underscore that there is a pressing need for future work to simultaneously consider both (Figures 3–8, Tables 4–7)—particularly when ε is systematic in nature (Section 4.2.1)—as it can potentially bias estimates of ET_{CWB} (Figures 5 and 6), which are widely used in upland settings (Barnhart et al., 2016; Berghuijs et al., 2014; Condon & Maxwell, 2017; Greve et al., 2020). Consistent with prior work on the influence of groundwater in the Budyko space (Hahm, Dralle et al., 2019; Hahm, Rempe et al., 2019; Istanbulluoglu et al., 2012; Jones et al., 2012), we observed that catchments tracked slightly below the Budyko curve. We observed slightly lower-than-expected ET fractions based on modeled ET_{Budyko} per Equation 12 when constrained to catchments with strong physical support for inferred groundwater behavior (see squares in Figure 7). However, large ε in the underlying water budget led to more uniform and dramatic deviations from ET on modeled ET_{Budyko} (see triangles and inverted triangles in Figure 7).

Understanding whether and how changes in precipitation phase will influence streamflow remains a pressing challenge as reflected in contrasting literature results (Barnhart et al., 2016; Berghuijs et al., 2014; Gordon et al., 2022; McCabe et al., 2018; Milly & Dunne, 2020). By imposing CWB assumptions in upland catchments, some research (Berghuijs et al., 2014; Ni et al., 2015) has used the Budyko hypothesis to posit that

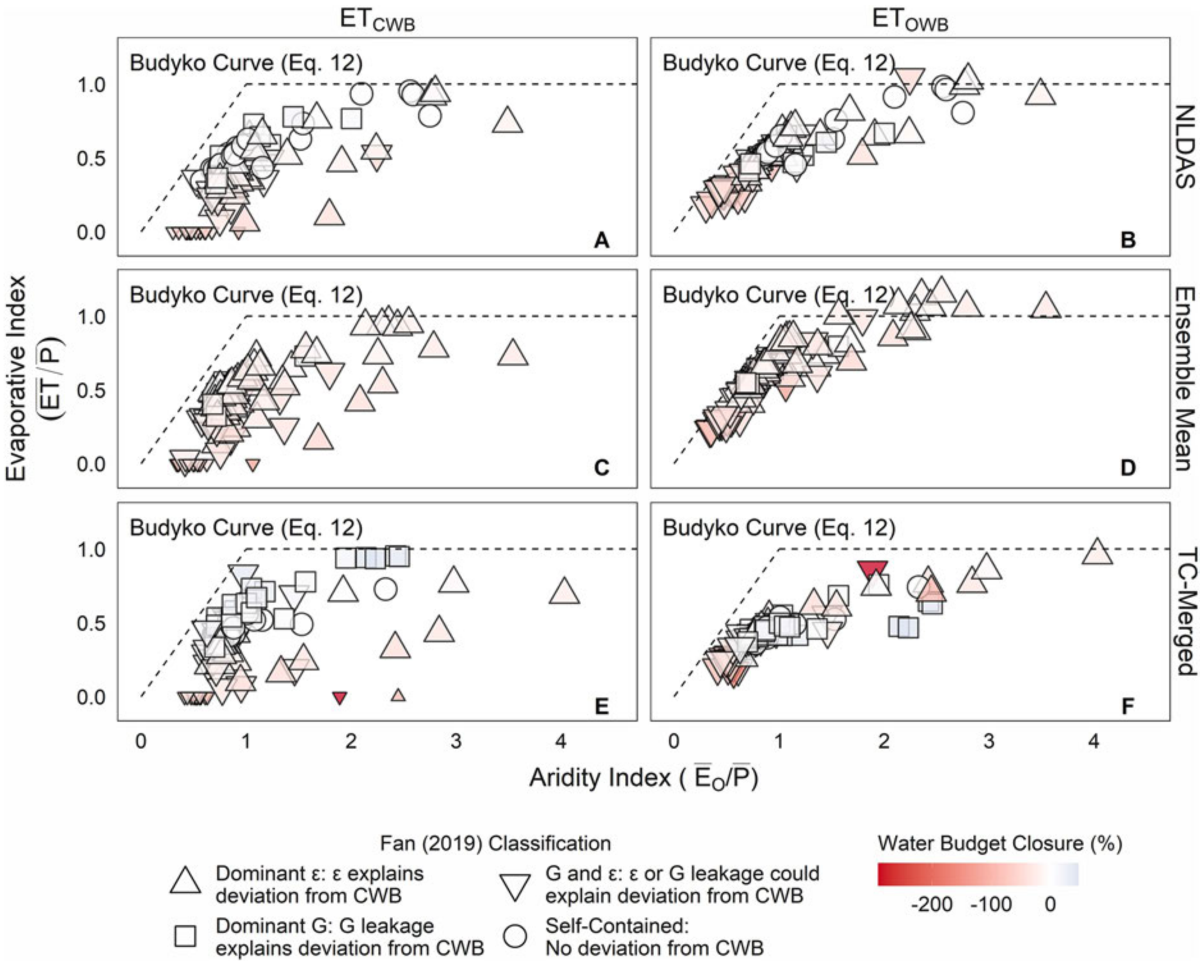

Figure 7. Variability in Budyko plots across all catchments (n = 114) using ET_{CWB} combined with: (a) NLDAS P; (c) Ensemble Mean P; and (e) TC-Merged P. Variability in Budyko plots across all catchments using ET_{OWB} combined with: (b) NLDAS P & ET; (d) Ensemble Mean P & ET; and (f) TC-Merged P & ET. Shapes correspond to Fan (2019) classification based on Figure 5. Coloring is based on observed OWB closure as calculated according to Section 3.2. Evaporative Index values less than 0 were forced to 0 for plotting purposes and are denoted as smaller symbols.

higher fs influences lower-than-expected ET fractions (higher-than-expected Q_{anom}) based on modeled ET_{Budyko}. However, snow-dominated upland catchments where fs is higher are also likely to experience non-zero G (Carroll et al., 2019) and are particularly susceptible to systematic under-prediction of P (Dettinger, Cayan et al., 2014; Dettinger, Redmond, & Cayan, 2004; He et al., 2019; Henn et al., 2018; Wrzesien et al., 2019). We also observed non-zero G and ε in the majority of our catchments (Table 4), the balance of which was propagated into the relationship between Q_{anom} and fs via ET_{CWB}(Figure 8, Table 7). Across all products, we found that the relationship between Q_{anom} and fs was highly sensitive to broken assumptions about ε when $ET_{\mathrm{CWB}}$was used. Using TC-Merged ET_{CWB}, the sensitivity of Q_{anom} to fs was 5.5 times greater in catchments impacted by ε than in catchments with physically supported G and 3 times greater than in self-contained catchments (Figure 8, Table 7). Within catchments impacted by ε, we further observed that the use of ET_{CWB} increased the sensitivity of Q_{anom} to fs by 9 times when compared to ET_{OWB}. Catchments impacted by ε tended to be steeper, snowier catchments in the western US (Figure 5, S9 in Supporting Information S1), highlighting that systematic under-prediction of P can bias Budyko-based inferences about upland water supplies in important and unconsidered ways.

Using TC-Merged P & ET in catchments with strong physical support for OWB closure (Section 4.2) also led to weaker relationships between Q_{anom} and fs than in previous results (slope = 0 13 for ET_{CWB}, slope = 0.11 for ET_{OWB} per Table 7 vs. slope = 0.37 from Berghuijs et al. (2014) and 0.32 from Barnhart et al. (2016)). These findings are consistent with recent research showing that climate change-driven declines in fs have not led to anticipated declines in streamflow efficiency in the Western US (McCabe et al., 2018) and/or that rain may countervail reductions in streamflow from declining fs (Hammond & Kampf, 2020). Given this, we suggest that caution is

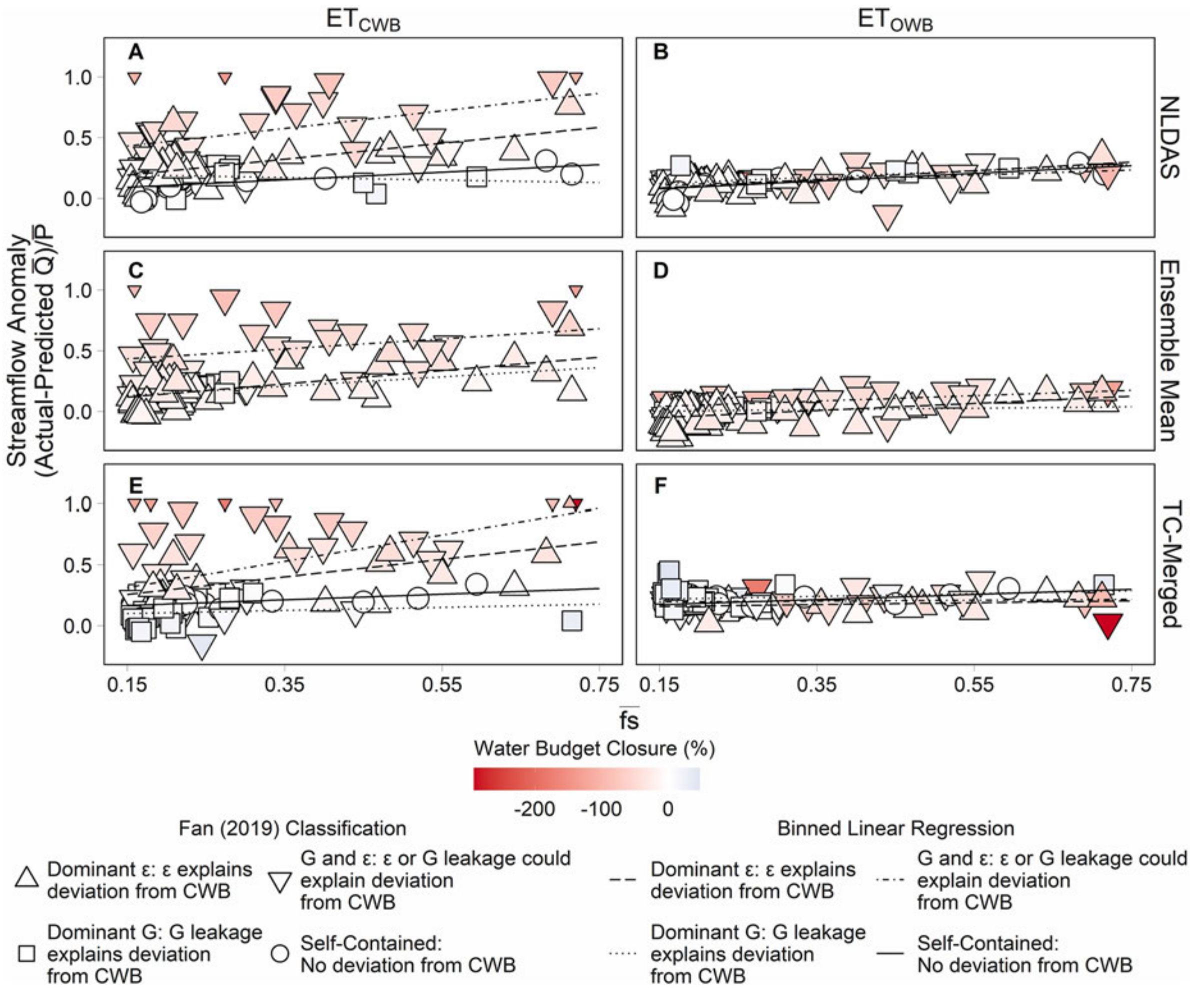

Figure 8. Variability in the relationship between Budyko streamflow anomaly and *fs* across all catchments ($n = 114$) using ET_{CWB} combined with: (a) NLDAS *P*; (c) Ensemble Mean *P*; and (e) TC-Merged *P*. Variability in the relationship between Budyko streamflow anomaly and *fs* using ET_{OWB} combined with: (b) NLDAS *P* & *ET*; (d) Ensemble Mean *P* & *ET*; and (f) TC-Merged *P* & *ET*. Shapes correspond to Fan (2019) classification based on Figure 5. Coloring is based on the evaluation of Equation (2) using ET_{OWB}. Line-type corresponds to binned linear regressions based on catchment grouping according to Fan (2019) classification based on Figure 5. Consistent with Figure 6, Q_{anom} values greater than 1 were forced to 1 for plotting purposes and are denoted as smaller symbols.

warranted when applying the conventional (CWB-based) form of the Budyko hypothesis to make predictions about water supplies in upland settings where disentangling the effects of physical (e.g., *G*) and non-physical (e.g., *ε*) factors remains a challenge. Expansion of the Budyko equation and subsequently Eqaution 13 to include *G* and *ε* variables following prior work by Istanbulluoglu et al. (2012) may be one way to enable more robust inferences.

5.3. When ε Is Characterized Using TC, Can the Fan (2019) Framework and ET_{OWB} Products Improve Insights About G Export or Import in Upland Settings?

CWBs and associated tools like the Budyko hypothesis ignore *G* to derive ET_{CWB} and make inferences about upland surface water supplies over longer timescales. However, evidence increasingly suggests that upland catchments can import/export substantial groundwater resources even over longer periods of time (Condon et al., 2020; Fan, 2019; Fan & Schaller, 2009). We found that TC-Merged *P* & *ET* facilitated insights about *G* that were not possible with other products tested (Figures 3–5) and importantly, that are ignored altogether in many applications of CWBs and CWB-based tools including the Budyko hypothesis (Sections 5.1 and 5.2). Consistent with expectations based on relevant literature (Fan & Schaller, 2009; Welch & Allen, 2012), TC-Merged *P* & *ET* substantially increased the number of catchments classified as groundwater exporters based on inferred groundwater behavior (i.e., positive $G + \varepsilon$ based on Equation 2 over other products (33 with TC-Merged *P* & *ET* vs. 7 with NLDAS *P* & *ET* vs. 0 with Ensemble Mean per Table 4). When inferred groundwater behavior

Table 7
Table of the Binned Linear Regression Equations and Correlation Coefficients Corresponding With Figures 7a–7f

		ET_{CWB}		ET_{OWB}	
Data product	Number of Fan (2019) criteria met	Equation	r	Equation	r
NLDAS	Self-Contained	$y = 0.31x + 0.04$	0.66	$y = 0.32x + 0.03$	0.68
	Dominant ε	$y = 0.66x + 0.09$	0.62	$y = 0.35x + 0.03$	0.63
	Dominant ε or G	$y = 0.65x + 0.38$	0.32	$y = 0.21x + 0.08$	0.44
	Dominant G	$y = -0.10x + 0.21$	−0.15	$y = 0.26x + 0.08$	0.60
Ensemble mean	Self-contained	–	–	–	–
	Dominant ε	$y = 0.55x + 0.04$	0.560	$y = 0.29x–0.09$	0.53
	Dominant ε or G	$y = 0.81x + 0.27$	0.44	$y = 0.22x + 0.01$	0.46
	Dominant G	$y = 0.38x + 0.07$	0.27	$y = 0.06x–0.01$	0.11
TC-Merged	Self-contained	$y = 0.23x + 0.13$	0.58	$y = 0.21x + 0.14$	0.62
	Dominant ε	$y = 0.72x + 0.15$	0.58	$y = 0.08x + 0.15$	0.28
	Dominant ε or G	$y = 1.68x + 0.14$	0.41	$y = 0.04x + 0.2$	0.09
	Dominant G	$y = 0.13x + 0.08$	0.13	$y = 0.11x + 0.20$	0.14

Note. Here, $y = \overline{Q_{\text{anom}}}$ and $x = \overline{fs}$.

was further evaluated using the criteria outlined in Section 3.2., TC-Merged P & ET yielded 33 catchments with strong physical support for inferred G (Table 6), which was almost two times the catchments NLDAS P & ET or Ensemble Mean P & ET yielded (19 catchments each, respectively). Overall, statistically characterizing random ε via TC substantially increased the number of catchments with physically supported G based on five testable physical criteria drawn from Fan (2019) (Figure 5, Table 6). Furthermore, results using TC-Merged P & ET confirmed the profile of groundwater exporting catchments drawn by Fan (2019): that they are positioned higher up in their containing sub-basins, are drier, and have deeper permeable regolith or fractured rock (Figures 4 and 5, Tables 5 and 6). When combined with Sections 5.1 and 5.2, this suggests that there is robust potential for novel combinations of TC-based merging and the Fan (2019) framework to harness advances in independent estimates of ET_{OWB} to improve predictions about critical surface and groundwater resources originating in upland settings.

5.4. Limitations and Paths Forward

Accurately closing the water budget in data-limited upland settings has eluded hydrologists for decades (Kampf et al., 2020) and is likely to remain elusive for many more (Safeeq et al., 2021). However, incremental progress that weaves together novel tools like TC and physically based frameworks like the one presented by Fan (2019) into OWBs can facilitate critical insight about upland water supplies. Nevertheless, several limitations are worthy of further discussion.

First, near term measurement limitations in interbasin groundwater fluxes and stores (Condon & Maxwell, 2017; Fan, 2019) undeniably exist and were a limitation in this study. Due to these limitations, we were restricted in our validation of inferred groundwater behavior and neglected the contribution of $\frac{\Delta s}{\Delta t}$ due to the too coarseness of existing data (Tapley et al., 2004) and challenges associated with distinguishing G from $\frac{\Delta s}{\Delta t}$ (Enzminger et al., 2019). However, a number of promising paths are being pursued to estimate groundwater fluxes and stores using gravity based measurements such as the Gravity Recovery and Climate Experiment (GRACE) mission (Tapley et al., 2004) and follow-on mission (Flechtner et al., 2014). Here too, there is potential value in microwave-based soil moisture retrievals (Crow et al., 2017), which could serve as a down-scaling tool. Other approaches, such as hydro-geophysical characterization (Gordon et al., 2020; Schmidt & Rempe, 2020; Smith et al., 2017), or some combination of the above in combination with the water budget (Hahm, Dralle et al., 2019; Hahm, Rempe et al., 2019) may also be helpful in improving our understanding of upland groundwater resources. At larger scales, however, physically based frameworks like the one put forward by Fan (2019) can help to improve and refine our understanding of the characteristics that promote interbasin groundwater import or export beyond those illustrated here. There are, however, challenges associated with linking qualitative criteria to quantitative data in a robust and repeatable manner. In this study, we used a Kruskal-Wallis and pairwise Wilcoxon rank sum to first establish statistical relationships between water budget closure and continuous data used to approximate the physical criteria proposed by Fan (2019). We then use conditional inference trees to establish rules relating observed water budget closure to physical criteria. Although our results indicate strong statistical support for the use of Fan (2019), future work could improve linkages between this generalizable framework and continuous data. When incorporated into the template put forward by Istanbulluoglu et al. (2012), advances in the quality and availability of groundwater data could assist simple tools like the Budyko hypothesis in more robustly accounting for conventionally ignored water budget variables.

Second, the use of TC in our study reduced the number of study catchments from 268 to 114, which underscores an important limitation of this method, at least based on the levels of correlated errors in currently available products. Because our methodology required valid TC outputs (e.g., outputs that do not violate assumptions listed in Section 3.1) for both P and ET, it is not surprising that many catchments had to be excluded in our analysis. Interestingly, valid ET estimates proved to be more challenging for TC ($n = 143$ valid estimates) than valid P outputs ($n = 170$). These challenges could be attributed in part to the pervasiveness of systematic errors in the underlying P data, including persistent under-catch, which tend to be more resilient to filtering via averaging (Yilmaz & Crow, 2014). Despite these challenges, TC enabled physically supported insight about G and ε that were otherwise obscured using other products (Figures 3–8). Improvements in bias-correction methodologies and underlying estimations (Beck et al., 2020) could work in concert with advances in data and modeling to expand the number of products with independent error sources available in upland settings. Of particular interest is the role dynamic atmospheric models like the Weather Research and Forecasting (WRF) model may have in improving estimates of P by better accounting for topographical features and mountain-precipitation interactions (He et al., 2019; Lundquist et al., 2019). Advances in remote sensing may also improve detection of solid and/or mixed phase P in complex topography (Lundquist et al., 2008; Maggioni et al., 2016), which is a goal of the on-going Global Precipitation Measurement (GPM) mission (Skofronick-Jackson et al., 2018). Continued improvement in measurement technology can help further eliminate systematic errors, enhancing the efficiency of TC for statistically characterizing and filtering random error in upland water budgets.

6. Conclusions

Water budgets and associated tools like the Budyko hypothesis are likely to remain central to investigations of upland catchment behavior. Recent advances in approaches to condition expectations about G (e.g., Fan (2019)), tools to characterize ε (e.g., TC), and products to estimate ET_{OWB} offer an opportunity for users to pivot away from the restrictive—and increasingly fragile—assumptions required to impose water budget closure. Motivated by this opportunity, our study sought to better understand the value of these advances for understanding the validity and consequences—in terms of inferences about surface and groundwater —of CWB assumptions (i.e., $G = 0$, $\varepsilon = 0 \rightarrow ET_{\mathrm{CWB}} = P–Q$) in a range of upland settings.

Here we find that propagating largely unsupported CWB assumptions into ET_{CWB} had a profound effect on inferences about upland surface and groundwater resources when assessed against an OWB assisted by a physical framework proposed by Fan (2019), TC, and independent estimates of ET_{OWB}. In particular, we observed that unconsidered ε can unrealistically alter expectations about streamflow response to climate change and mask groundwater contributions. Long term OWB closure supported non-zero G and ε in 85 of 114 catchments (~75%) using TC-Merged P & ET, 89 of 114 catchments (~78%), using NLDAS P & ET, and all 114 catchments using Ensemble Mean P & ET over a 15-year period. When these were screened based using five testable criteria proposed by Fan (2019), TC-Merged P & ET led to physically realistic G in 33 of 114 catchments (~29%)—the most of any product tested. Using TC-Merged estimates of ET_{CWB}, we observed inflated (5.5 times higher) streamflow response to fs in catchments with large ε compared to those with smaller suspected ε. Furthermore, within catchments with large ε, we found that TC-Merged ET_{CWB} increased the sensitivity of streamflow to fs by 9 times more than TC-Merged ET_{OWB}. Because ET_{CWB} is commonly used to make these types of inferences, our results highlight the need for users to consider the effects of ε—in addition to physical factors such as G—in discussions about catchment behavior in the Budyko space. More fundamentally though, the widespread conflict between CWB assumptions and observed OWB closure in upland catchments points to the need for users to critically evaluate the Budyko assumptions when error and groundwater flow are expected to be high.

In the vital pursuit of improved predictions about upland water supplies, our results demonstrate the value of TC and a physically-based framework proposed by Fan (2019) for harnessing advances in the estimation of ET_{OWB} and expanding the use of OWBs. While CWB approaches—and to a lesser extent conventional products (e.g., a standard model and ensemble mean)—overlooked strong evidence for G, the combination of TC-based merging and the physically-based framework proposed by Fan (2019) revealed groundwater exportation in high positioned, arid catchments with deep substrates. Albeit challenging, larger and more comprehensive groundwater datasets in upland settings could help to further refine insights about G in Budyko-type analyses. As modeling, remote sensing, data assimilation, and bias correction techniques in upland settings advance in tandem, combining TC and evidence-based frameworks is a promising path forward to improve predictions about upland water supplies.

Data Availability Statement

The authors provide our TC-Merged data in Supporting Information S1. With the exception of ALEXI data (Anderson et al., 2011), the original data products used in this study are freely available to the public: streamflow, Daymet precipitation, NLDAS precipitation, catchment shapefiles, and attributes can be retrieved from https://ral.ucar.edu/solutions/products/camels (last access: 20 July 2021) (Addor et al., 2017; Newman et al., 2015). ERA5 precipitation data were accessed at https://cds.climate.copernicus.eu/cdsapp#!/home (last access: 1 February 2020) (Hersbach et al., 2020) via Google Earth Engine, Persiann-CDR precipitation data were accessed at https://doi.org/10.7289/V51V5BWQ (last access: 1 February 2020) (Ashouri et al., 2015; Sorooshian et al., 2014), NLDAS precipitation data were accessed via https://ral.ucar.edu/solutions/products/camels and https://doi.org/10.1029/2010EO340001 (last access: 22 July 2020) (Xia et al., 2012) via Google Earth Engine, Daymet precipitation data were accessed via https://ral.ucar.edu/solutions/products/camels and https://doi.org/10.3334/ORNLDAAC/1840 (last access: 22 July 2020) (Thornton et al., 2016) via Google Earth Engine, PRISM precipitation data were accessed at https://doi.org/10.1371/journal.pone.0141140 (last access: 9 September 2019). SSEBop evapotranspiration data were accessed at https://earlywarning.usgs.gov/fews (last access: 22 July 2020) (Senay et al., 2013), NCA-LDAS data were accessed at https://disc.gsfc.nasa.gov/ (last access: 10 July 2020) (Kumar et al., 2019), and MODIS16 evapotranspiration data were accessed at https://www.ntsg.umt.edu/project/modis/mod16.php (last access: 2 February 2020) (Mu et al., 2013) via Google Earth Engine. Potential Evapotranspiration data were accessed via https://ral.ucar.edu/solutions/products/camels and https://doi.org/10.1029/2010EO340001 (last access: 1 August 2020) (Xia et al., 2012). Streamflow data were accessed at https://ral.ucar.edu/solutions/products/camels and https://waterdata.usgs.gov/nwis/rt (last access: 20 June 2020) (USGS, 2016). Snow fraction data were accessed at https://ral.ucar.edu/solutions/products/camels (last access: 20 July 2021).

Acknowledgments
Support for this research is provided the USDA NIFA (Project # NEVW-2017-08812), NSF EAR #2012310, and the Lincoln Institute's Babbitt Dissertation Fellowship Program. The authors would also like to thank the three anonymous reviewers, Associate Editor, and Editor for their constructive and detailed feedback on this manuscript. Many thanks to Dr. Gabrielle Boisrame for her constructive and thoughtful feedback on this manuscript and guidance on statistics. The authors are also grateful for Dr. Christopher Hain at NASA for generously providing and assisting with the ALEXI data, Dr. Sebastian Krogh, Mari Webb, and Aidan Manning for constructive feedback on an earlier draft of this manuscript, and Gary Sterle for assistance with figures. The USDA ARS is an equal opportunity employer.

References

Abolafia-Rosenzweig, R., Pan, M., Zeng, J. L., & Livneh, B. (2021). Remotely sensed ensembles of the terrestrial water budget over major global river basins: An assessment of three closure techniques. *Remote Sensing of Environment*, *252*, 112191. https://doi.org/10.1016/J.RSE.2020.112191

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Alemohammad, S. H., McColl, K. A., Konings, A. G., Entekhabi, D., & Stoffelen, A. (2015). Characterization of precipitation product errors across the United States using multiplicative triple collocation. *Hydrology and Earth System Sciences*, *19*(8), 3489–3503. https://doi.org/10.5194/hess-19-3489-2015

Anderson, M. C., Kustas, W. P., Norman, J. M., Hain, C. R., Mecikalski, J. R., Schultz, L., et al. (2011). Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. *Hydrology and Earth System Sciences*, *15*(1), 223–239. https://doi.org/10.5194/hess-15-223-2011

Andréassian, V., & Perrin, C. (2012). On the ambiguous interpretation of the Turc-Budyko nondimensional graph. *Water Resources Research*, *48*(10), 2012WR012532. https://doi.org/10.1029/2012WR012532

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., et al. (2015). PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, *96*(1), 69–83. https://doi.org/10.1175/BAMS-D-13-00068.1

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R., & Dozier, J. (2006). Mountain hydrology of the western United States. *Water Resources Research*, *42*(8), 8432. https://doi.org/10.1029/2005WR004387

Barnett, T. P., Pierce, D. W., Hidalgo, H. G., Bonfils, C., Santer, B. D., Das, T., et al. (2008). Human-induced changes in the hydrology of the Western United States. *Science*, *319*(5866), 1080–1083. https://doi.org/10.1126/science.1152538

Barnhart, T. B., Molotch, N. P., Livneh, B., Harpold, A. A., Knowles, J. F., & Schneider, D. (2016). Snowmelt rate dictates streamflow. *Geophysical Research Letters*, *43*(15), 8006–8016. https://doi.org/10.1002/2016GL069690

Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O. M., et al. (2020). Bias correction of global high-resolution precipitation climatologies using streamflow observations from 9372 catchments. *Journal of Climate*, *33*(4), 1299–1315. https://doi.org/10.1175/JCLI-D-19-0332.1

Berghuijs, W. R., Gnann, S. J., & Woods, R. A. (2020). Unanswered questions on the Budyko framework. *Hydrological Processes*, *34*(26), 13958–15703. https://doi.org/10.1002/hyp.13958

Berghuijs, W. R., Woods, R. A., & Hrachowitz, M. (2014). A precipitation shift from snow towards rain leads to a decrease in streamflow. *Nature Climate Change*, *4*(7), 583–586. https://doi.org/10.1038/nclimate2246

Budyko, M. I. (1974). *Climate and life*. Academic press.

Burnett, M. W., Quetin, G. R., & Konings, A. G. (2020). Data-driven estimates of evapotranspiration and its controls in the Congo Basin. *Hydrology and Earth System Sciences*, *24*(8), 4189–4211. https://doi.org/10.5194/hess-24-4189-2020

Buss, H. L., Brantley, S. L., Scatena, F. N., Bazilievskaya, E. A., Blum, A., Schulz, M., et al. (2013). Probing the deep critical zone beneath the luquillo experimental forest, Puerto Rico. *Earth Surface Processes and Landforms*, *38*(10), 1170–1186. https://doi.org/10.1002/esp.3409

Carroll, R. W. H., Deems, J. S., Niswonger, R., Schumer, R., & Williams, K. H. (2019). The importance of interflow to groundwater recharge in a snowmelt-dominated headwater basin. *Geophysical Research Letters*, *46*(11), 5899–5908. https://doi.org/10.1029/2019GL082447

Chen, F., Crow, W. T., Bindlish, R., Colliander, A., Burgin, M. S., Asanuma, J., & Aida, K. (2018). Global-scale evaluation of SMAP, SMOS, and ASCAT soil moisture products using triple collocation. *Remote Sensing of Environment*, *214*, 1–13. https://doi.org/10.1016/j.rse.2018.05.008

Condon, L. E., & Maxwell, R. M. (2017). Systematic shifts in Budyko relationships caused by groundwater storage changes. *Hydrology and Earth System Sciences*, *21*(2), 1117–1135. https://doi.org/10.5194/hess-21-1117-2017

Condon, L. E., Markovich, K. H., Kelleher, C. A., McDonnell, J. J., Ferguson, G., & McIntosh, J. C. (2020). *Where is the bottom of a watershed? Water resources research*. Blackwell Publishing Ltd. https://doi.org/10.1029/2019WR026010

Crow, W. T., Han, E., Ryu, D., Hain, C. R., & Anderson, M. C. (2017). Estimating annual water storage variations in medium-scale (2000-10000 km^2) basins using microwave-based soil moisture retrievals. *Hydrology and Earth System Sciences*, *21*(3), 1849–1862. https://doi.org/10.5194/HESS-21-1849-2017

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., et al. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *28*(15), 2031–2064. https://doi.org/10.1002/joc.1688

Daly, C., Taylor, G., & Gibson, W. (1997). The prism approach to mapping precipitation and temperature.

Dettinger, M. D., Cayan, D. R., Meyer, M. K., & Jeton, A. (2004). Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American River basins, Sierra Nevada, California, 1900-2099. *Climatic Change*, *62*(1–3), 283–317. https://doi.org/10.1023/B:CLIM.0000013683.13346.4f

Dettinger, M., Redmond, K., & Cayan, D. (2004). Winter orographic precipitation ratios in the Sierra Nevada–Large-scale atmospheric circulations and hydrologic consequences. *Journal of Hydrometeorology*, *5*(6), 1102–1116. https://doi.org/10.1175/JHM-390.1

Dong, J., Crow, W. T., Duan, Z., Wei, L., & Lu, Y. (2019). A double instrumental variable method for geophysical product error estimation. *Remote Sensing of Environment*, *225*, 217–228. https://doi.org/10.1016/j.rse.2019.03.003

Ehsani, N., Vörösmarty, C. J., Fekete, B. M., & Stakhiv, E. Z. (2017). Reservoir operations under climate change: Storage capacity options to mitigate risk. *Journal of Hydrology*, *555*, 435–446. https://doi.org/10.1016/j.jhydrol.2017.09.008

Enzminger, T. L., Small, E. E., & Borsa, A. A. (2019). Subsurface water dominates Sierra Nevada seasonal hydrologic storage. *Geophysical Research Letters*, *46*(21), 11993–12001. https://doi.org/10.1029/2019GL084589

Fan, Y. (2019). Are catchments leaky? *WIREs Water*, *6*(6). https://doi.org/10.1002/wat2.1386

Fan, Y., & Schaller, M. F. (2009). River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. *Journal of Geophysical Research*, *114*, D04103. https://doi.org/10.1029/2008JD010636

Flechtner, F., Morton, P., Watkins, M., & Webb, F. (2014). Status of the GRACE follow-on mission. *International Association of Geodesy Symposia*, *141*, 117–121. https://doi.org/10.1007/978-3-319-10837-7_15/COVER

Frisbee, M. D., Tysor, E. H., Stewart-Maddox, N. S., Tsinnajinnie, L. M., Wilson, J. L., Granger, D. E., & Newman, B. D. (2016). Is there a geomorphic expression of interbasin groundwater flow in watersheds? Interactions between interbasin groundwater flow, springs, streams, and geomorphology. *Geophysical Research Letters*, *43*(3), 1158–1165. https://doi.org/10.1002/2015GL067082

Genereux, D. P., Nagy, L. A., Osburn, C. L., & Oberbauer, S. F. (2013). A connection to deep groundwater alters ecosystem carbon fluxes and budgets: Example from a Costa Rican rainforest. *Geophysical Research Letters*, *40*(10), 2066–2070. https://doi.org/10.1002/grl.50423

Gleeson, T., Moosdorf, N., Hartmann, J., & Van Beek, L. P. H. (2014). A glimpse beneath Earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. *Geophysical Research Letters*, *41*(11), 3891–3898. https://doi.org/10.1002/2014GL059856

Gordon, B. L., Brooks, P. D., Krogh, S. A., Boisrame, G. F. S., Carroll, R. W. H., McNamara, J. P., & Harpold, A. A. (2022). Why does snowmelt-driven streamflow response to warming vary? A data-driven review and predictive framework. *Environmental Research Letters*, *17*(5), 053004. https://doi.org/10.1088/1748-9326/AC64B4

Gordon, B. L., Paige, G. B., Miller, S. N., Claes, N., & Parsekian, A. D. (2020). Field scale quantification indicates potential for variability in return flows from flood irrigation in the high altitude western US. *Agricultural Water Management*, *232*, 106062. https://doi.org/10.1016/j.agwat.2020.106062

Greve, P., Burek, P., & Wada, Y. (2020). Using the Budyko framework for calibrating a global hydrological model. *Water Resources Research*, *56*(6), e2019WR026280. https://doi.org/10.1029/2019WR026280

Gruber, A., Dorigo, W. A., Crow, W., & Wagner, W. (2017). Triple collocation-based merging of satellite soil moisture retrievals. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(12), 6780–6792. https://doi.org/10.1109/TGRS.2017.2734070

Gruber, A., Su, C.-H., Zwieback, S., Crow, W., Dorigo, W., & Wagner, W. (2016). Recent advances in (soil moisture) triple collocation analysis. *International Journal of Applied Earth Observation and Geoinformation*, *45*, 200–211. https://doi.org/10.1016/j.jag.2015.09.002

Hahm, W. J., Dralle, D. N., Rempe, D. M., Bryk, A. B., Thompson, S. E., Dawson, T. E., & Dietrich, W. E. (2019). Low subsurface water storage capacity relative to annual rainfall decouples Mediterranean plant productivity and water use from rainfall variability. *Geophysical Research Letters*, *46*(12), 6544–6553. https://doi.org/10.1029/2019GL083294

Hahm, W. J., Rempe, D. M., Dralle, D. N., Dawson, T. E., Lovill, S. M., Bryk, A. B., et al. (2019). Lithologically controlled subsurface critical zone thickness and water storage capacity determine regional plant community composition. *Water Resources Research*, *55*(4), 3028–3055. https://doi.org/10.1029/2018WR023760

Hamilton, A. S., & Moore, R. D. (2012). Quantifying uncertainty in streamflow records. *Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques*, *37*(1), 3–21. https://doi.org/10.4296/cwrj3701865

Hammond, J. C., & Kampf, S. K. (2020). Subannual streamflow responses to rainfall and snowmelt inputs in snow-dominated watersheds of the western United States. *Water Resources Research*, *56*(4), e2019WR026132. https://doi.org/10.1029/2019WR026132

Han, P. F., Istanbulluoglu, E., Wan, L., & Wang, X. S. (2021). A new hydrologic sensitivity framework for unsteady-state responses to climate change and its application to catchments with croplands in Illinois. *Water Resources Research*, *57*(8), e2020WR027762. https://doi.org/10.1029/2020WR027762

Harpold, A., Brooks, P., Rajagopal, S., Heidbuchel, I., Jardine, A., & Stielstra, C. (2012). Changes in snowpack accumulation and ablation in the intermountain west. *Water Resources Research*, *48*(11), W11501. https://doi.org/10.1029/2012WR011949

He, C., Chen, F., Barlage, M., Liu, C., Newman, A., Tang, W., et al. (2019). Can convection-permitting modeling provide decent precipitation for offline high-resolution snowpack simulations over mountains? *Journal of Geophysical Research: Atmospheres*, *124*, 12631–12654. https://doi.org/10.1029/2019JD030823

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hothorn, T., Hornik, K., & Zeileis, A. (2012). Unbiased recursive partitioning: A conditional inference framework. *The American Statistician*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, T., Hornik, K., Van De Wiel, M. A., & Zeileis, A. (2012). A lego system for conditional inference. *The American Statistician*, *60*(3), 257–263. https://doi.org/10.1198/000313006X118430

Hothorn, T., Zeileis, A., Cheng, E., & Ong, S. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909.

Immerzeel, W. W., Lutz, A. F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., et al. (2020). Importance and vulnerability of the world's water towers. *Nature*, *577*(7790), 364–369. https://doi.org/10.1038/s41586-019-1822-y

Istanbulluoglu, E., Wang, T., Wright, O. M., & Lenters, J. D. (2012). Interpretation of hydrologic trends from a water balance perspective: The role of groundwater storage in the Budyko hypothesis. *Water Resources Research*, *48*(3), W00H16. https://doi.org/10.1029/2010WR010100

Jing, X., Geerts, B., Wang, Y., & Liu, C. (2017). Evaluating seasonal orographic precipitation in the interior western United States using gauge data, gridded precipitation estimates, and a regional climate simulation. *Journal of Hydrometeorology*, *18*(9), 2541–2558. https://doi.org/10.1175/JHM-D-17-0056.1

Jones, J. A., Creed, I. F., Hatcher, K. L., Warren, R. J., Adams, M. B., Benson, M. H., et al. (2012). Ecosystem processes and human influences regulate streamflow response to climate change at long-term ecological research sites. *BioScience*, *62*(4), 390–404. https://doi.org/10.1525/BIO.2012.62.4.10

Kampf, S. K., Burges, S. J., Hammond, J. C., Bhaskar, A., Covino, T. P., Eurich, A., et al. (2020). The case for an open water balance: Re-envisioning network design and data analysis for a complex, uncertain world. *Water Resources Research*, *56*(6), e2019WR026699. https://doi.org/10.1029/2019WR026699

Khan, M. S., Liaqat, U. W., Baik, J., & Choi, M. (2018). Stand-alone uncertainty characterization of GLEAM, GLDAS, and MOD16 evapotranspiration products using an extended triple collocation approach. *Agricultural and Forest Meteorology*, *252*, 256–268. https://doi.org/10.1016/j.agrformet.2018.01.022

Koppa, A., & Gebremichael, M. (2017). A Framework for validation of remotely sensed precipitation and evapotranspiration based on the Budyko hypothesis. *Water Resources Research*, *53*(10), 8487–8499. https://doi.org/10.1002/2017WR020593

Kumar, S. V., Jasinski, M., Mocko, D. M., Rodell, M., Borak, J., Li, B., et al. (2019). NCA-LDAS land analysis: Development and performance of a multisensor, multivariate land data assimilation system for the national climate assessment. *Journal of Hydrometeorology*, *20*(8), 1571–1593. https://doi.org/10.1175/JHM-D-17-0125.1

Käser, D., & Hunkeler, D. (2016). Contribution of alluvial groundwater to the outflow of mountainous catchments. *Water Resources Research*, *52*(2), 680–697. https://doi.org/10.1002/2014WR016730

Li, D., Pan, M., Cong, Z., Zhang, L., & Wood, E. (2013). Vegetation control on water and energy balance within the Budyko framework. *Water Resources Research*, *49*(2), 969–976. https://doi.org/10.1002/wrcr.20107

Li, D., Wrzesien, M. L., Durand, M., Adam, J., & Lettenmaier, D. P. (2017). How much runoff originates as snow in the western United States, and how will that change in the future? *Geophysical Research Letters*, *44*(12), 6163–6172. https://doi.org/10.1002/2017GL073551

Lundquist, J. D., Neiman, P. J., Martner, B., White, A. B., Gottas, D. J., & Ralph, F. M. (2008). Rain versus Snow in the Sierra Nevada, California: Comparing Doppler profiling radar and surface observations of melting level. *Journal of Hydrometeorology*, *9*(2), 194–211. https://doi.org/10.1175/2007JHM853.1

Lundquist, J., Dickerson-Lange, S., Gutmann, E., Jonas, T., Lumbrazo, C., & Reynolds, D. (2021). Snow interception modeling: Isolated Observations have led to land surface models lacking appropriate climate sensitivities. *Authorea Preprints*. https://doi.org/10.22541/AU.161443094.42414018/V1

Lundquist, J., Hughes, M., Gutmann, E., & Kapnick, S. (2019). Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bulletin of the American Meteorological Society*, *100*(12), 2473–2490. https://doi.org/10.1175/BAMS-D-19-0001.1

Maggioni, V., Sapiano, M. R. P., & Adler, R. F. (2016). Estimating uncertainties in high-resolution satellite precipitation products: Systematic or random error? *Journal of Hydrometeorology*, *17*(4), 1119–1129. https://doi.org/10.1175/JHM-D-15-0094.1

Mankin, J. S., Viviroli, D., Singh, D., Hoekstra, A. Y., & Diffenbaugh, N. S. (2015). The potential for snow to supply human water demand in the present and future. *Environmental Research Letters*, *10*(11), 114016. https://doi.org/10.1088/1748-9326/10/11/114016

Massari, C., Crow, W., & Brocca, L. (2017). An assessment of the performance of global rainfall estimates without ground-based observations. *Hydrology and Earth System Sciences*, *21*(9), 4347–4361. https://doi.org/10.5194/hess-21-4347-2017

Maxwell, R. M., & Condon, L. E. (2016). Connections between groundwater flow and transpiration partitioning. *Science*, *353*(6297), 377–380. https://doi.org/10.1126/science.aaf7891

McCabe, G. J., Wolock, D. M., & Valentin, M. (2018). Warming is driving decreases in snow fractions while runoff efficiency remains mostly unchanged in snow-covered areas of the western United States. *Journal of Hydrometeorology*, *19*(5), 803–814. https://doi.org/10.1175/JHM-D-17-0227.1

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters*, *41*(17), 6229–6236. https://doi.org/10.1002/2014GL061322

Milly, P. C. D., Dunne, K. A., Yu, Z., Kong, P., Sefik, E., Xu, H., et al. (2020). Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation. *Science*, *367*(6483), 1252–1255. https://doi.org/10.1126/science.aax0194

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

Mu, Q., Zhao, M., & Running, S. W. (2013). MODIS global terrestrial evapotranspiration (ET) product.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Ni, G., Yang, D., Zhang, D., Cong, Z., Ni, G., Yang, D., & Hu, S. (2015). Effects of snow ratio on annual runoff within the Budyko framework. *Hydrology and Earth System Sciences*, *19*(4), 1977–1992. https://doi.org/10.5194/hess-19-1977-2015

Padrón, R. S., Gudmundsson, L., Greve, P., & Seneviratne, S. I. (2017). Large-Scale controls of the surface water balance over land: Insights from a systematic review and meta-analysis. *Water Resources Research*, *53*(11), 9659–9678. https://doi.org/10.1002/2017WR021215

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G. Y., et al. (2016). A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *Journal of Advances in Modeling Earth Systems*, *8*(1), 41–65. https://doi.org/10.1002/2015MS000526

Polhamus, A., Fisher, J. B., & Tu, K. P. (2013). What controls the error structure in evapotranspiration models? *Agricultural and Forest Meteorology*, *169*, 12–24. https://doi.org/10.1016/j.agrformet.2012.10.002

Potter, N. J., Zhang, L., Milly, P. C. D., McMahon, T. A., & Jakeman, A. J. (2005). Effects of rainfall seasonality and soil moisture capacity on mean annual water balance for Australian catchments. *Water Resources Research*, *41*(6), 1–11. https://doi.org/10.1029/2004WR003697

Qin, Y., Abatzoglou, J. T., Siebert, S., Huning, L. S., AghaKouchak, A., Mankin, J. S., et al. (2020). Agricultural risks from changing snowmelt. *Nature Climate Change*, *10*(5), 459–465. https://doi.org/10.1038/s41558-020-0746-8

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., et al. (2012). How well are we measuring snow: The NOAA/FAA/NCAR winter precipitation test bed. *Bulletin of the American Meteorological Society*, *93*(6), 811–829. https://doi.org/10.1175/BAMS-D-11-00052.1

Safeeq, M., Bart, R. R., Pelak, N. F., Singh, C. K., Dralle, D. N., Hartsough, P., & Wagenbrenner, J. W. (2021). How realistic are water-balance closure assumptions? A demonstration from the southern sierra critical zone observatory and Kings River experimental watersheds. *Hydrological Processes*, *35*(5), e14199. https://doi.org/10.1002/hyp.14199

Schmidt, L., & Rempe, D. (2020). Quantifying dynamic water storage in unsaturated bedrock with borehole nuclear magnetic resonance. *Geophysical Research Letters*, *47*(22), e2020GL089600. https://doi.org/10.1029/2020GL089600

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

Skofronick-Jackson, G., Kirschbaum, D., Petersen, W., Huffman, G., Kidd, C., Stocker, E., & Kakar, R. (2018). The Global Precipitation Measurement (GPM) mission's scientific achievements and societal contributions: Reviewing four years of advanced rain and snow observations. *Quarterly Journal of the Royal Meteorological Society*, *144*(S1), 27–48. https://doi.org/10.1002/qj.3313

Smith, R. G., Knight, R., Chen, J., Reeves, J. A., Zebker, H. A., Farr, T., & Liu, Z. (2017). Estimating the permanent loss of groundwater storage in the southern San Joaquin Valley, California. *Water Resources Research*, *53*(3), 2133–2148. https://doi.org/10.1002/2016WR019861

Sorooshian, S., Hsu, K., Braithwaite, D., Ashouri, H., & Program, N. C. (2014). PERSIANN-CDR. https://doi.org/10.7289/V51V5BWQ

Sposito, G. (2017). Understanding the Budyko equation. *Water*, *9*(4), 236. https://doi.org/10.3390/w9040236

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple. *Collocation*, *103*(3334), 7755–7766. https://doi.org/10.1029/97jc03180

Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F., & Watkins, M. M. (2004). GRACE measurements of mass variability in the Earth system. *Science*, *305*(5683), 503–505. https://doi.org/10.1126/science.1099192

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wei, Y., Devarakonda, R., Vose, R. S., & Cook, R. B. (2016). Daymet: Daily surface weather data on a 1-km grid for North America, version 3. ORNL DAAC, Oak Ridge, Tennessee, USA. *USDA-NASS, 2019. 2017 Census of Agriculture, Summary and State Data, Geographic Area Series, Part*, *51*.

U.S. Geological Survey. (2016). *National water information system data available on the World Wide Web*. USGS Water Data for the Nation). Retrieved from https://waterdata.usgs.gov/nwis/

Valéry, A., Andréassian, V., & Perrin, C. (2010). Regionalization of precipitation and air temperature over high-altitude catchments–learning from outliers. *Hydrological Sciences Journal*, *55*(6), 928–940. https://doi.org/10.1080/02626667.2010.504676

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

Viviroli, D., Dürr, H. H., Messerli, B., Meybeck, M., & Weingartner, R. (2007). Mountains of the world, water towers for humanity: Typology, mapping, and global significance. *Water Resources Research*, *43*(7), 7447. https://doi.org/10.1029/2006WR005653

Voepel, H., Ruddell, B., Schumer, R., Troch, P. A., Brooks, P. D., Neal, A., et al. (2011). Quantifying the role of climate and landscape characteristics on hydrologic partitioning and vegetation response. *Water Resources Research*, *47*(10).

Wang, T., Istanbulluoglu, E., Lenters, J., Scott, D., Wang, T., Istanbulluoglu, E., et al. (2009). On the role of groundwater and soil texture in the regional water balance: An investigation of the Nebraska Sand Hills, USA. *Water Resources Research*, *45*(10), W10413. https://doi.org/10.1029/2009WR007733

Welch, L. A., & Allen, D. M. (2012). Consistency of groundwater flow patterns in mountainous topography: Implications for valley bottom water replenishment and for defining groundwater flow boundaries. *Water Resources Research*, *48*(5), W05526. https://doi.org/10.1029/2011WR010901

Wrzesien, M. L., Durand, M. T., & Pavelsky, T. M. (2019). A reassessment of North American River basin cool-season precipitation: Developments from a new mountain climatology data set. *Water Resources Research*, *55*(4), 3502–3519. https://doi.org/10.1029/2018WR024106

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*, 3109. https://doi.org/10.1029/2011JD016048

Yang, L., Meng, X., & Zhang, X. (2011). SRTM DEM and its application advances. *International Journal of Remote Sensing*, *32*(14), 3875–3896. https://doi.org/10.1080/01431161003786016

Yilmaz, M. T., & Crow, W. T. (2014). Evaluation of assumptions in soil moisture triple collocation analysis. *Journal of Hydrometeorology*, *15*(3), 1293–1302. https://doi.org/10.1175/JHM-D-13-0158.1

Yilmaz, M. T., Crow, W. T., Anderson, M. C., & Hain, C. (2012). An objective methodology for merging satellite- and model-based soil moisture products. *Water Resources Research*, *48*(11), W11502. https://doi.org/10.1029/2011WR011682

References From the Supporting Information

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., et al. (2003). Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research*, *108*, 2002JD003118. https://doi.org/10.1029/2002jd003118

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., et al. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research*, *108*, 2002JD003296. https://doi.org/10.1029/2002jd003296

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Higgins, R. W., Schemm, J. K. E., Shi, W., & Leetmaa, A. (2000). Extreme precipitation events in the Western United States related to tropical forcing. *Journal of Climate*, *13*(4), 793–820. https://doi.org/10.1175/1520-0442(2000)013<0793:EPEITW>2.0.CO;2

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, *116*, D12109. https://doi.org/10.1029/2010JD015139

Peng, L., Li, D., & Sheffield, J. (2018). Drivers of variability in atmospheric Evaporative demand: Multiscale spectral analysis based on observations and physically based modeling. *Water Resources Research*, *54*(5), 3510–3529. https://doi.org/10.1029/2017WR022104

Rui, H., & Mocko, D. (2018). README for NCA-LDAS version 2.0 data product 2.

Thornton, P. E., Running, S. W., & White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, *190*(3–4), 214–251. https://doi.org/10.1016/S0022-1694(96)03128-9

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