AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR018748

Improving Budyko curve-based estimates of long-term water partitioning using hydrologic signatures from GRACE

Kuai Fang[1], Chaopeng Shen[1], Joshua B. Fisher[2], and Jie Niu[3,4]

[1]Department of Civil and Environmental Engineering, Pennsylvania State University, State College, Pennsylvania, USA, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA, [3]Lawrence Berkeley National Laboratory, Berkeley, California, USA, [4]University of California, Santa Barbara, Santa Barbara, California, USA

Key Points:
- GRACE amplitude can predict departure of water partitioning from the Budyko curve
- GRACE-based models have better transferability than physical-factor-based models
- Our global product is comparable to other ET products and can be useful in data fusion

Correspondence to:
C. Shen,
cshen@engr.psu.edu

Citation:
Fang, K., C. Shen, J. B. Fisher, and J. Niu (2016), Improving Budyko curve-based estimates of long-term water partitioning using hydrologic signatures from GRACE, *Water Resour. Res.*, *52*, 5537–5554, doi:10.1002/2016WR018748.

Received 9 FEB 2016
Accepted 31 MAY 2016
Accepted article online 2 JUN 2016
Published online 30 JUL 2016

Abstract The Budyko hypothesis provides a first-order estimate of water partitioning into runoff (Q) and evapotranspiration (E). Observations, however, often show significant departures from the Budyko curve; moreover, past improvements to Budyko curve tend to lose predictive power when migrated between regions or to small scales. Here to estimate departures from the Budyko curve, we use hydrologic signatures extracted from Gravity Recovery And Climate Experiment (GRACE) terrestrial water storage anomalies. The signatures include GRACE amplitude as a fraction of precipitation (A/P), interannual variability, and 1-month lag autocorrelation. We created a group of linear models embodying two alternate hypotheses that departures can be predicted by (a) Taylor series expansion based on the deviation of physical characteristics (seasonality, snow fraction, and vegetation index) from reference conditions and (b) surrogate indicators covarying with E, e.g., A/P. These models are fitted using a mesoscale USA data set (HUC4) and then evaluated using world data sets and USA basins $<1 \times 10^5$ km^2. The model with A/P could reduce error by 50% compared to Budyko itself. We found that seasonality and fraction of precipitation as snow account for a major portion of the predictive power of A/P, while the remainder is attributed to unexplained basin characteristics. When migrated to a global data set, type b models performed better than type a. This contrast in transferability is argued to be due to data set limitations and catchment coevolution. The GRACE-based correction performs well for USA basins >1000 km^2 and, according to comparison with other global data sets, is suitable for data fusion purposes, with GRACE error as estimates of uncertainty.

1. Introduction

The Budyko hypothesis [*Budyko*, 1948; *Arora*, 2002; *Gerrits et al*., 2009; *Wang and Tang*, 2014] describes the long-term partitioning of precipitation (P) between evapotranspiration (E) and runoff (Q), as a function of the ratio between potential evapotranspiration (E_p) and P, also called the aridity index (E_p/P), i.e.,

$$\frac{E}{P} = f\left(\frac{E_p}{P}\right), \tag{1}$$

where E/P is termed the evaporation ratio, and f stands for the Budyko curve, for which many formulations exist, e.g., the Turk-Pike equation [*Pike*, 1964; *Yang et al*., 2008] modified by *Chen et al*. [2013] with the addition of an abscissa-intercept term

$$f\left(\frac{E_p}{P}\right) = \left[1 + \left(\frac{E_p}{P} - \varphi\right)^{-2}\right]^{-1/2}, \tag{2}$$

where φ is the intercept added by *Chen et al*. [2013]. Recently, the Budyko curve has found wide applications as a *reference* condition [*Istanbulluoglu et al*., 2012; *Berghuijs et al*., 2014a; *Carmona et al*., 2014] or providing a framework for the understanding of hydrologic controls [*Gentine et al*., 2012] under climate change [*Berghuijs et al*., 2014a; *Yang et al*., 2014; *Zhang et al*., 2015]. On a theoretical level, the Budyko curve initiated an interesting hypothesis that various catchment characteristics coevolved with climate to manifest such a simple water partitioning pattern [*Troch et al*., 2013; *Li et al*., 2014]. On a practical level, the Budyko curve provides an independent first-order estimate of E for predictions in ungauged basins (PUB) [*Hrachowitz et al*., 2013], without the use of any hydrologic models, which is useful for the evaluation of land surface

© 2016. American Geophysical Union. All Rights Reserved.

hydrologic models [e.g., *Xia et al*., 2012; *Oleson et al*., 2013; *Shen et al*., 2013, 2014, 2016; *Clark et al*., 2015; *Fatichi et al*., 2016].

However, recent studies have focused on noticeable deviations from the traditional theoretical Budyko curve, i.e., there can often be a significant departure

$$\frac{E}{P}=f\left(\frac{E_p}{P}\right)+\delta, \tag{3}$$

where $\delta=\frac{P-Q}{P}-f\left(\frac{E_p}{P}\right)$ is the departure of actual evaporation ratio (*E/P*) from the Budyko-predicted value (for long-term water balance, *E* can be well approximated by $P-Q$). First, attention was paid to climatic pattern, in particular, the phase difference between *E* and precipitation in an arid region (aridity index >1) [*Chen et al*., 2013; *Berghuijs et al*., 2014b]. If precipitation and E_p peaks are "in-sync" throughout a year, nearly all precipitation will become *E* and E/P will be close to 1; in contrast, if *P* concentrates in the winter when there is little E_p, actual *E* will be significantly less than that of the uniform case. Besides climatic factors, researchers also examined influence of physical characteristics, especially vegetation control. *Li et al*. [2013] used the normalized difference vegetation index (*NDVI*) to parameterize Fu's version of the Budyko equation [*Fu*, 1981; *Zhang et al*., 2001] and found that vegetation control is more apparent for larger basins ($>$300,000 km^2) but diminishes for basins smaller than 50,000 km^2. *Xu et al*. [2013] parameterized Fu's equation at different scales with *NDVI*, topography, latitude, longitude, and elevation. Their calibrated coefficients are also different for basins of different sizes. The performance degradation and the required changes in coefficients for small scales found in the above studies need to be better understood. For prediction in ungauged basins, special attention is needed for the generality of the method to avoid overfitting. A generalized formula that is portable across scales and regions can also help advance process-level understanding of water partitioning.

Although remote sensing methods have made large strides recently, there is not yet direct measurement of *E*. Satellite-based products for *E*, e.g., MOD16A2 [*Mu et al*., 2011] rely on assumptions and empirical formulations. The GRACE mission [*Tapley et al*., 2004; *Wahr*, 2004] records terrestrial water storage anomalies (TWSA, storage deviation from the long-term mean) for the world and has been shown to be useful for a variety of applications including monitoring groundwater resources [*Famiglietti et al*., 2011; *Scanlon et al*., 2012; *Döll et al*., 2014; *Huang et al*., 2015], model calibration and testing [*Lo et al*., 2010; *Niu et al*., 2014], flood forecasting [*Reager et al*., 2014] and drought monitoring [*Long et al*., 2014a]. Since storage is a competing process of runoff and *E*, we expect TWSA to contain signals relevant to runoff and *E*. For example, the amplitude of the TWSA (*A*, average peak height from the mean) as fraction of *P* is an indicator of the relative strengths of storage and release of the system (climate and catchment). Relevant to *E*, the use of GRACE is ordinarily in a data assimilation/model calibration setting [e.g., *Long et al*., 2014b]. No effort, to the authors' best knowledge, examines how GRACE TWSA is related to *E* in a Budyko framework.

Hydrologic "signatures" [*Vogel and Sankarasubramanian*, 2003; *Gupta et al*., 2008; *Yilmaz et al*., 2008] are statistics extracted from hydrologic time series to highlight certain distinguishing behaviors of the hydrologic systems. Streamflow-derived hydrologic signatures have been used in model calibration [*Yilmaz et al*., 2008], parameter regionalization [*Yadav et al*., 2007], and catchments classification [*Sawicz et al*., 2014]. In this paper, we attempt to answer the following questions: Does there exist a relationship between basin *E* and GRACE signatures, such as *A* as a fraction of *P*, which is useful for improving estimates of water partitioning, and, if so, what factors contribute to the *A-E* relationship? Is such a relationship general enough to be portable across different regions? At what scale is the equation valid? In the following, we first describe data sources, processing procedures, signatures and indicators computed, and the Analysis of Covariance (ANCOVA) used to partition the predictive power of *A/P* to different factors. Then we introduce the linear models for the departure term and their different underlying hypotheses. After that, we show the performance of the models across scales and regions and its control factors through variance partitioning. Finally, a new long-term global average *E* data set with error estimates is introduced. This data set is a new "hydrologic-model-free" independent validation data sets useful in data fusion.

2. Methods

2.1. Data Sources and Processing Procedures

We employed three sets of basins (Figures B1 in Appendix B) in order to comprehensively examine model portability and scale-dependence issues: (1) 179 Hydrologic Cataloging Unit 4-digit (HUC4) basins, which

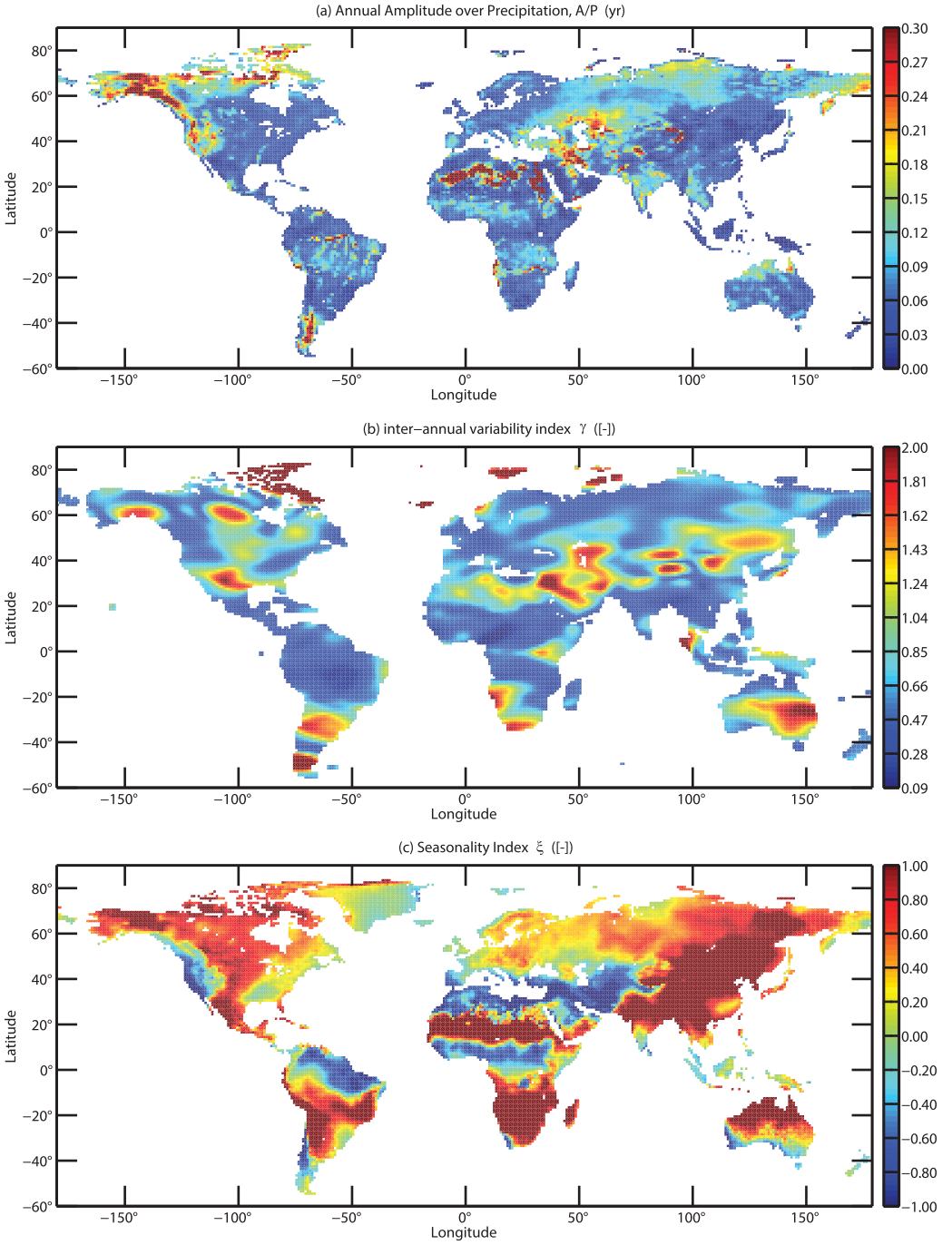

Figure 1. Hydrologic signatures and climatic index computed for the world using data from October 2002 to September 2010: (a) average annual GRACE TWSA amplitude as a fraction of precipitation (A/P); (b) the interannual GRACE signal variability ratio γ (dimensionless ratio between between-year and within-year TWSA variability); and (c) the precipitation-temperature seasonality index following Woods [2009], which is −1 for completely out of phase, 1 for completely in phase, and 0 for uncorrelated.

seamlessly cover the conterminous USA; (2) 605 basins from the global runoff data center (GRDC) that are within 10^4-10^5 km^2; and (3) 4627 US Geological Survey (USGS) gauged basins from the GAGES-II data set [*Falcone*, 2011] which has been used to analyze climate change imprint on alluvial rivers [*Slater and Singer*, 2013] and hydraulic geometry [*Shen et al.*, 2016]. We screened the data for temporal coverage (sites with less than 90% for the period from 1 October 2002 to 31 December 2012 are removed) and spatial coverage (sites with catchment areas bigger than the HUC4 they are located in are removed, because they are downstream gages in major rivers).

The three sets of basins have different forcing data sources. For the HUC4 basins, hourly precipitation, temperature, radiation, wind speeds, and humidity from the North American Land Data Assimilation System (NLDAS) [*Xia et al*., 2012] were aggregated to the basins for the calculation of E_p using the Shuttleworth equation (section 2.3). Monthly runoff was obtained from USGS website (http://waterwatch.usgs.gov/new/index.php?id=romap3). This version of HUC4 runoff was computed by aggregating flow from data-sufficient gages located within a HUC4. For the USGS GAGES-II basins, NLDAS climate forcing was distributed into the basin boundaries, similar to the HUC4 data set. Daily discharges from USGS websites were downloaded from USGS website (http://waterdata.usgs.gov/nwis/) and aggregated to analysis time periods/scales. We compared HUC4 data with intersecting MOPEX data sets [*Duan et al*., 2006] and the corresponding P and E_p are similar (Figure B2).

For the GRDC basins, long-term annual average discharge was obtained from the GRDC data set (http://www.bafg.de/GRDC/EN/Home/homepage_node.html). Precipitation was derived from two data sets. For tropical and midlatitude regions (between latitudes -50° and $+50^\circ$), we employed the precipitation product from the Tropical Rainfall Measuring Mission (TRMM) [*Huffman et al*., 1997] (3B42V7 derived). For high-latitude basins, we used precipitation forcing data from the Global Land Data Assimilation System (GLDAS) version 2 [*Rodell et al*., 2004]. The version 2 GLDAS product is only available until the end of 2010. Other climatic inputs are extracted from GLDAS for all basins.

Following *Li et al*. [2013], we obtained the 10 km resolution *NDVI* from Global Inventory Modeling and Mapping Studies (GIMMS) [*Buermann*, 2002]. We used the average *NDVI* for the period 1984–2006 in our analysis. For E comparison, we used the PT-JPL product described in *Fisher et al*. [2008], hereafter termed E^{PJ}. This approach utilizes ecophysiological constraint functions to downscale E_p to actual E using remotely sensed observations of land and atmosphere properties. The algorithm and product have been widely used and independently validated extensively throughout the scientific literature, showing top performance across multiple intercomparisons [e.g., *Vinukollu et al*., 2011; *Chen et al*., 2014; *McCabe et al*., 2015; *Miralles et al*., 2015]. In addition, we also compared our E estimate with simulated E from GLDAS version 2 with the NOAH land surface model [*Ek et al*., 2003].

Currently, three GRACE solutions [*Swenson*, 2014] are provided at monthly time intervals and 1-degree spatial resolution for the world. GRACE TWSA mass grids level 3 version 5.0 data, processed using University of Texas Center for Space Research (CSR) algorithm was downloaded from GRACE Tellus website [*Swenson*, 2012]. The GRACE product uses a destriping filter and a 300 km wide Gaussian filter as well to minimize North-South stripes in the monthly maps. The scaling factor based on land surface models, proposed by *Swenson and Wahr* [2006] and *Landerer and Swenson* [2012], was applied to the original gridded data to restore signal losses due to surface mass variations at small spatial scales tend to be attenuated by the low-pass filtering of GRACE spherical harmonics. GRACE TWSA data and hydrologic signatures are averaged by area to the HUC4 and GRDC basins. For the USGS basins, since some of them are too small, they are assigned the GRACE data from the HUC4 in which they are located.

2.2. Hydrologic Signatures Extracted From GRACE

We extracted three hydrologic signatures from GRACE monthly time series (October 2002 and September 2014), based on reasoning about the hydrologic systems and statistical significance: (1) the average annual maximum TWSA amplitude (A) as a fraction of precipitation (A/P) (Figure 1a). This signature is chosen because the fraction of precipitation stored reflects the competition between storage, runoff and ET. Water stored through infiltration or snowpack accumulation is more likely to be released as runoff. Therefore, we anticipate that higher A/P is correlated with lower E/P. More explanations are provided in section 2.4; (2) the ratio of TWSA variance explained by interannual variability and intra-annual variability (γ) (Figure 1b). γ is chosen because basins with higher interannual variability in storage (normalized to intra-annual variability) tend to have smaller long-term average E/P compared more evenly distributed ones: in years with extraordinary precipitation, there is a higher chance of water partitioned as storage and runoff as opposed to E. Interannual variability was also found to be important for long-term water balance [*Sivapalan et al*., 2011; *Li*, 2014]; and (3) acf_D, the 1 month lag, piecewise-detrended autocorrelation function of GRACE TWSA based on D-month-long segments. acf_D describes the smoothness of the monthly TWSA signal and reflects the seasonal distribution pattern of storage. As with γ, a higher concentration of storage in a few months will likely result in smaller E/P compared to average conditions. Apart from climate pattern, longer memory in

storage may indicate the ability of the system to hold water from runoff, hypothetically leading to higher *E/P* compared to low-memory systems.

A is estimated by first applying a Fourier-transform to the GRACE time series, and then take the maximum amplitude for frequencies between 0.8 and 1.2 cycles/yr. Previous research showed that most mega-basins in the world have the highest peaks at annual periods [*Reager and Famiglietti*, 2013]. Although in that work the Yule-Walker autoregressive method [*Emery and Thomson*, 2004] was used, we are only interested in the annual amplitude and therefore a Fourier transform is sufficient. The band window of 0.8–1.2 is for numerical stability of the method and slight changes of the window did not change our results. We can see hot spots of *A/P* in areas near large rivers, e.g., Mississippi, lower Nile, and Amazon (Figure 1a). As described previously, in these regions *A/P* is not reflective of land surface runoff and storage, but rather, seasonal river stage fluctuations, so they are removed from model fitting. When working with HUC4 data set, since the Mississippi River induce large seasonal mass changes and leakage errors which are unrelated to nearby land surface runoff/water storage dynamics, the basins that contain the Mississippi River are removed from analysis. In the USA (a bigger map of *A/P* for HUC4 basins is presented in Figure B1 in the Appendix), the high *A/P* regions are in the west, where precipitation is winter-dominant [*Berghuijs et al*., 2014b] and has a big phase difference with temperature.

In addition, we propose an interannual variability index, γ (Figure 1b), which quantifies the ratio of between-year variability and within-year variability in TWSA. If $\sigma_{w,i}(S)$ is the standard deviation of TWSA in *i*-th year (based on monthly data), the average within-year standard deviation is $\overline{\sigma_w}=\frac{1}{n_y}\sum_{i=1}^{n_y}\sigma_{w,i}(TWSA)$, where n_y is the number of years. The mean of each year's TWSA is $\overline{TWSA}_i$, and σ_b is the standard deviation of $\overline{TWSA}_i$. We define γ as

$$\gamma=\frac{\sigma_b}{\overline{\sigma_w}}. \tag{4}$$

We calculated the indices using GRACE data from October 2002 to September 2014. Some data gaps in GRACE has been filled using spline interpolation. We evaluated using data from 2002 to 2012 which has fewer gaps. This was not found to have significant influence on our results.

acf_D is the cross correlation of the GRACE data with itself with 1-month lag, after applying piecewise detrending in every *D* months. Higher acf_D indicates higher similarity between data points and their neighbors and thus higher smoothness. Lower acf_D curves have more abrupt changes. We first divided the time series into multiple segments, each consisting of *D* months of data (in our analysis we used $D = 48$). Without piecewise detrending, nonstationary trends may interfere with the extraction of the autocorrelation function. It is well-known that different time periods can exhibit different trends in the GRACE data [see, e.g., *Famiglietti et al*., 2011; *Voss et al*., 2013]. Here to simplify the analysis, we only examined 1-lag *acf* with 48 months as the segment length for detrending. We tested 2-month lag and 3-month lag (data not shown here), which did not provide much additional predictive power.

GRACE data are influenced by different sources of errors [*Wahr et al*., 2006]. Signal degradation due to measurement noises are called measurement errors [*Swenson and Wahr*, 2006] and the contamination of signal by nearby region (due to spectral truncation and filtering) is termed leakage errors (Figure B3). We employed measurement and leakage errors estimated using the approach in Landerer and Swenson [2012] and provided at http://grace.jpl.nasa.gov to calculate the combined error as the quadratic mean of the two errors. The errors are used to determine regions where GRACE-based signatures have low reliability, which are overwritten by interpolation. After our initial testing, we found that where the combined error is large (>74 mm, which is two times the global mean combined error), the hydrologic signatures from GRACE is no longer usable. For world data sets (GRDC and world-gridded products), regions with errors larger than 74 mm obtain their *A/P* via interpolation from other regions. In fact, few GRDC basins fall into this category. *A/P* for cells with major rivers such as Amazon, lower Nile, and Mississippi are automatically interpolated from neighboring regions.

2.3. Climatic Indices

E_p was calculated using the Shuttleworth equation [*Shuttleworth*, 1993; *Zhou et al*., 2006; *Li et al*., 2013]

$$\lambda E_p=\frac{\Delta\cdot R_n}{\Delta+\gamma_p}+\frac{6.43\gamma_p}{\Delta+\gamma_p}\times(1+0.5361u)\times\frac{(e_s-e_a)}{C_p}, \tag{5}$$

where Δ is the rate of change of saturation specific humidity (kPa $°C^{-1}$), γ_p is Psychrometric constant (kPa $°C^{-1}$), R_n is net radiation (MJ m^{-2} d^{-1}), λ is the latent heat of vaporization (MJ kg^{-1}), u is wind speed

(m s^{-1}), C_p is specific heat of evaporation (MJ kg^{-1} $°C^{-1}$), e_s is saturation vapor pressure (kPa), e_a is near-surface air vapor pressure (kPa), and the unit of E_p from this equation is mm d^{-1}. The daily E_p was aggregated to mm yr^{-1} for the calculation of the aridity index.

We calculated the fraction of *P* as snow (*S/P*), which is known to be important for runoff [*Berghuijs et al*., 2014a]. We also calculated a seasonality index (Figure 1c) that quantifies the phase difference between precipitation and temperature following previous work [*Woods*, 2009; *Berghuijs et al*., 2014b], who derived it based on fitting data to the following equations which adopt sine curve assumptions [*Milly*, 1994; *Potter et al*., 2005]:

$$P(t)=\overline{P}\left[1+\Delta_p \sin\left(2\pi\left(t-s_p\right)/\tau\right)\right],$$

$$T(t)=\overline{T}+\Delta_t \sin\left(2\pi(t-s_t)/\tau\right),$$

$$\xi=\Delta_p \times sign(\Delta_t)\times \cos\left(2\pi\left(s_p-s_t\right)/\tau\right), \tag{6}$$

where t is the time (months), s_p and s_t are a phase shift for precipitation and temperature, respectively (months), τ is the duration of the seasonal cycle (12 months) and Δ_p and Δ_t are the dimensionless seasonal amplitudes for precipitation and temperature, respectively.

In theory, the seasonality index varies within $[-1,\ 1]$ and is -1 for completely out-of phase E_p and *P*, 0 for uniform precipitation with seasonal temperature and 1 for completely in phase E_p and *P*. However, in practice because some basins have very dry seasons, the fitted sine curves can have >1 amplitudes. Note that while ξ is negatively correlated with the *A/P* map (e.g., the western USA has high *A/P* and most negative ξ), the correlation is not perfect (e.g., the northern central lowland has relatively large *A/P* and also relatively large ξ).

2.4. The General Departure Model and Its Rationale

Our linear formula for the departure term is

$$\delta^*(\boldsymbol{x})=\delta(\boldsymbol{x})+\varepsilon=\boldsymbol{a}^T\boldsymbol{x}, \tag{7}$$

where δ^* is an approximation to δ, $\boldsymbol{x}$ is a vector of independent physical or surrogate factors (except 1 is the first element for the intercept term), or a mixture of both, $\boldsymbol{a}$ are the corresponding linear coefficients and ε is the error term.

Although equation (7) seems a simple linear regression model, it in fact embodies different hypotheses with different predictors. Here we make the distinction between physical factors (ξ, γ, *S/P*, and *NDVI*), which are independent variables that vary in space, and surrogate factors (*A/P*, acf_{48}, and γ), which are dependent variables potentially influenced by the former. When equation (7) involves only physical factors, it is an approximation to the relationships between potentially causal factors and outcome (departure from Budyko). Further, it can be interpreted as a first-order Taylor Series expansion of the perturbation from the reference state (Budyko curve) due to changes in physical factors (Appendix A). However, when a surrogate factor is involved in equation (7), there is no causal or controlling relationship: *E/P* does not change *because of* changes in *A/P*. Thus, a Taylor Series expansion interpretation is inappropriate. Rather, the surrogate factors and *E/P* covary due to changes in some common factors, and equation (7) captures their covariation. When we test the different models we are also testing different hypotheses, i.e., whether the departure from Budyko is better modeled by the difference in a list of physical factors or covarying surrogate indices. More mathematical discussion of the different hypotheses is provided in Appendix A.

There are multiple reasons behind choosing A/P. The most important one is the following observation: if the climate is such that *P* and E_p reach peaks and lows at approximately the same time, e.g., in the US central high plains, they are "in-phase." For "in-phase" and water-limited basins, *P* immediately evaporates, leaving very little water for storage and runoff and thus high *E/P*, but at the same time, the A/P ratio is also small because little water can be stored. On the other hand, if *P* and E_p are "out-of-phase," as in the case of US southwest, *P* during winter times "evades" the peak of E_p and has the chance to be stored. The storage of water in winter times leads to higher *A/P*, and at the same time a low E/P. Therefore, there should be negative covariation between the δ and A/P.

There are also many other potential "negative-A/P-E/P-correlation-inducing" (NAECI) physical factors. Whether a factor is NAECI depends on how, when it is varied, it shifts the competition between *E*, *Q*, and *S*. NAECI factors, in general, should favor infiltration against evaporation, for example, high vertical soil hydraulic conductivity in deep-water-table regions: more water infiltrate below plant-accessible zone, increasing groundwater storage while reducing *E*. Conversely, if a region has low vertical conductivity, it inhibits groundwater storage while enhancing *E*, still contributing to a negative *A/P-E/P* correlation. For another example, consistently light rainfall and cloudy patterns can promote infiltration and inhibit *E*. Some other processes may in fact cause positive *A/P* and *E/P* correlations. For example, high terrain slope potentially boosts runoff while decreasing both storage and evaporation, thus causing a positive correlation between the two. A factor's influence on *A/P-E/P* correlation can also be complicated and climate-dependent. For example, unusually high vegetation cover is an NAECI factor for an arid catchment as it boosts *E* and reduces infiltration (and thus storage); however, for low-aridity, energy-limited basins, vegetation interception primarily modulates runoff and storage, with little impact on *E*.

2.5. Partitioning of Variance

We used analysis of covariance (ANCOVA) to attribute the variability of δ to various physical factors or surrogate indices. Compared to the Analysis of Variance (ANOVA), which is designed for categorical data, ANCOVA builds linear models for continuous variables to attribute variance to predictors [*Keppel and Wickens*, 2004]. When the data are unbalanced (uneven sampling of data in different parts of the viable ranges of factors), as is the case with our data, there are three different ways of attributing the variance (sum of squares, SS): type I (sequential), type II (main effects excluded), and type III (main and interaction effect excluded) [*Fox*, 2008]. With type I, SS are attributed to the factors in the order of what is supplied, and only residuals are attributed to the next factors. Thus, the earlier factors will claim part of the SS of subsequent, correlated factors. For example, if two factors are perfectly correlated, the first factor in the sequence will be attributed all the SS they can explain and the second will be attributed none. With type II SS, the main effects of other factors are excluded so the only part of SS that can be solely attributed to one factor is reported. As a result, type II SS is not influenced by the order of factors. Type III is similar to type II, but further excluded interaction terms. Here we examined type I and type II to see how different variables overlap. In our ANCOVA we examined a total of six factors: *A/P*, ξ, γ, *S/P*, *NDVI*, and acf_{48}, and this order is called order 1 (O1). In the second order (O2), *A/P* is placed as the last argument: ξ, γ, *S/P*, *NDVI*, acf_{48}, and *A/P*.

2.6. Multiscale, Multi–Data Set Validation of Alternative Models

Since an important goal of the proposed method is to estimate *E* for ungauged basins, how the formula performs when coefficients are migrated between regions and across scales is of great importance. We tested the performance of a total of 13 linear models (Table 1), each with a different combination of the above-listed predictors, when their coefficients are estimated from the HUC4 data set and then migrated to the GRDC and USGS basins. The root mean squared error (RMSE) of *E/P* was used as a measure of error of the models

$$RMSE = \left\{ \frac{\sum_{j=1}^{n_b} \left[\left(f\left(E_p/P\right) + \delta^* - (P-Q)/P \right) \right]^2}{n_b} \right\}^{1/2}, \tag{8}$$

where n_b is the number of basins and δ^* is the model-predicted departure term. When validating the models using the USGS data set, we rank the basins in descending order by their areas. Based on this ranking the basins were evenly divided into 28 area classes. A separate RMSE was calculated for each class. Each plotted point is the average of two neighboring classes.

3. Results and Discussion

3.1. The GRACE-Assisted Departure Model

From the HUC4 data set, we note many basins, especially those with larger aridity index, deviate significantly from the Budyko curve (Figure 2a). For some arid basins, the ratio of Budyko-predicted and actual water partitioning, $f\left(\frac{E_p}{P}\right) : \frac{(P-Q)}{P}$, can be as large as 50–100%. As expected, these points are always accompanied by large *A/P*, showing a strong influence from seasonality. For wetter basins, on the other hand, there may be negative or positive departures, which may be small but could potentially lead to large errors in the

Table 1. Tested Linear Models With Linear Weights Provided[a]

Model #	Predictors
0	Budyko itself
1 (HUC4)	$0.224-1.884\ A/P$
1 (GRDC)	$0.219-1.293\ A/P$
2 (HUC4)	$0.216-1.824\ A/P+0.013\ \xi$
2 (GRDC)	$0.225-1.308\ A/P-0.017\ \xi$
3 (HUC4)	$0.316-1.871\ A/P+0.058\ \xi-0.128\ \gamma$
3 (GRDC)	$0.195-1.301\ A/P-0.021\ \xi+0.044\ \gamma$
4 (HUC4)	$0.327-1.691\ A/P+0.068\ \xi-0.142\ \gamma\ -0.182\ S/P$
4 (GRDC)	$0.200-1.061\ A/P-0.005\ \xi+0.056\ \gamma\ -0.262\ S/P$
5 (HUC4)	$0.170-1.351\ A/P+0.126\ \xi-0.127\ \gamma\ +0.054\ S/P\ +0.306\ NDVI$
5 (GRDC)	$0.128-1.052\ A/P-0.004\ \xi+\ 0.060\ \gamma-0.147\ S/P\ +0.161 NDVI$
6 (HUC4)	$-0.023-1.350\ A/P+0.127\ \xi-0.123\ \gamma\ +0.052\ S/P+0.336\ NDVI+0.223\ acf_{48}$
6 (GRDC)	$-0.503-1.195\ A/P+0.010\ \xi+0.091\ \gamma\ -0.103\ S/P+0.147\ NDVI+0.782\ acf_{48}$
7 (HUC4)	$0.180-1.776\ A/P+0.132\ NDVI$
7 (GRDC)	$0.134-1.143\ A/P+0.232\ NDVI$
8 (HUC4)	$0.330-2.098\ A/P-0.106\ \gamma$
8 (GRDC)	$0.191-1.284\ A/P+0.039\ \gamma$
9 (HUC4)	$0.313-2.101\ A/P-0.106\ \gamma+0.021\ acf_{48}$
9 (GRDC)	$-0.542-1.407\ A/P+0.084\ \gamma+0.905\ acf_{48}$
10 (HUC4)	$0.031+0.155\ \xi$
10 (GRDC)	$0.092+0.000\ \xi$
11 (HUC4)	$0.116+0.198\ \xi-0.114\ \gamma$
11 (GRDC)	$0.056-0.005\ \xi+0.053\ \gamma$
12 (HUC4)	$-0.034+0.251\ \xi-0.132\ \gamma-0.031\ S/P+\ 0.589\ NDVI$
12 (GRDC)	$0.029+0.018\ \xi+0.073\ \gamma-0.297\ S/P+\ 0.173\ NDVI$
13 (HUC4)	$-0.231+0.252\ \xi-0.127\ \gamma-0.033\ S/P+\ 0.619\ NDVI+0.228\ acf_{48}$
13 (GRDC)	$-0.400+0.029\ \xi+0.095\ \gamma-0.281\ S/P+\ 0.165\ NDVI+0.522\ acf_{48}$

[a]HUC4 means the model was fitted using HUC4 data. GRDC means the model was fitted using GRDC data. Model #9 was chosen to produce a global-scale comparison with E^{PJ} and E^{GLDAS}.

estimation of *E* due to the large *P*. We can clearly see there is a negative covariation between δ and *A/P* (Figure 2c), which is exploited by the predictive formula. After we subtract the corrector term, based solely on *A/P*, the points now cluster much more closely to the Budyko curve (Figure 2b).

3.2. Factors Contributing to Budyko Departures and *A/P-E/P* Correlation

The ANCOVA results show that for the HUC4 data set, *Precipitation* seasonality (ξ) and *S/P* are important but not the sole factors contributing to *A/P* and the departure. The sequential (type I) sum of squares (SS) attributed to *A/P*, when factors are laid out in O1, is more than 61% of the total SS, leaving only about ~3% to ξ (Figure 3). When they are laid in O2, in which *A/P* is placed last, both ξ and *S/P* explain much more variance than in O1, but there is still around 14% of SS attributed to *A/P* that is unexplained by any other factor. Therefore, a large fraction $\left(\frac{61\%-14\%}{61\%}=77\%\right)$ of the explanatory power of *A/P* is attributed to correlation with ξ and *S/P*. This pattern suggests that *P*-E_p phase shifts and snowpack accumulation are two major reasons *causing A/P* in the USA. In previous sections we reason that when *P* and *Ep* are "out-of-phase," rainwater has more opportunity to infiltrate, rather than becoming *E*, so that *E/P* is small compared to Budyko prediction while *A/P* is larger than average, giving rise to the negative correlation. *S/P* has a similar effect: when a larger fraction of precipitation falls as snow, snowpack accumulation produces larger *A/P*, while the snowmelt water becomes runoff or infiltration more easily compared to average conditions. In addition, there is little independent explanatory power in ξ and *S/P* that cannot be replaced by *A/P*, as reflected in type I SS in O1. This means *A/P* is an excellent surrogate index to represent their aggregate effects on long-term water partitioning. On another note, the interannual storage variability index has little correlation with *A/P* or ξ, as evidenced by the small change in its attributed SS from O1 to O2. When there is large interannual variability in precipitation, rainfall in wet years can create much higher runoff than average years and thus causes overall negative δ.

There are myriad processes that interact in complex ways to influence *A/P-E/P* correlation and the departure from Budyko, and they are not easily described by a small number of indices. However, at the scales of basins in the US and the world, such effects seem to be muted, and we observe primarily a negative *A/P-E/P* correlation, suggesting their influence on *E* is limited. Overall, *A/P* appears to be an effective way of capturing the lumped effects of these myriad processes, while the residual effects are in the error terms.

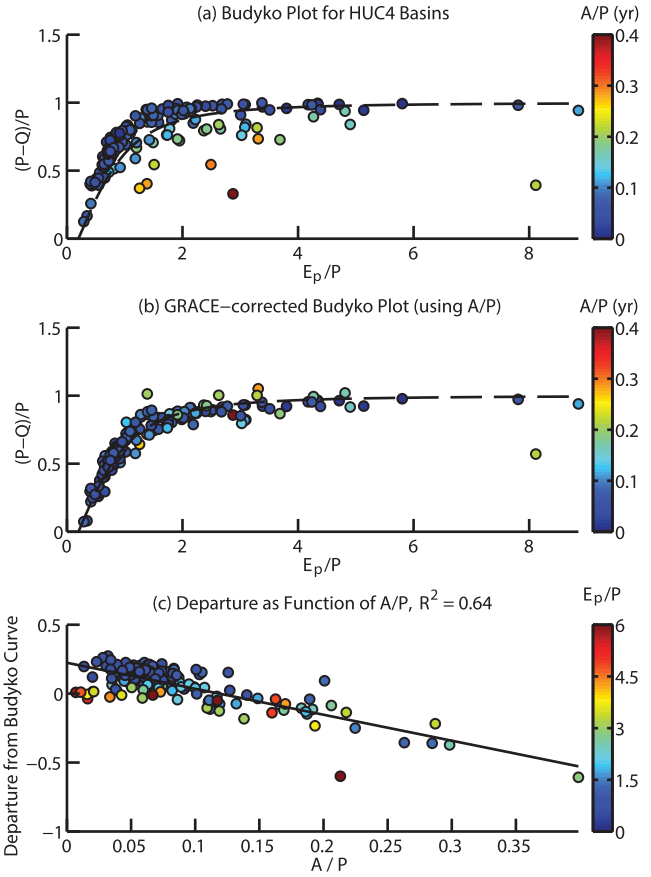

Figure 2. Using GRACE TWSA amplitude as a fraction of precipitation (A/P) to predict the departure (δ) from the Budyko curve for the HUC4 data set. $\delta = \frac{P-Q}{P} - f\left(\frac{E_p}{P}\right)$ where Q is observed discharge, Ep is potential evapotranspiration, and $f\left(\frac{E_p}{P}\right)$ is the Budyko formula. (a) Without correction, the HUC4 basins scatter around the Budyko curve, some with significant departures. (b) After correcting using A/P, basins are now much more closely clustered around the Budyko curve. (c) The negative correlation between A/P versus δ allows the improvement over Budyko.

3.3. Global Validation and the Relevance of Coevolution to Model Choosing

The HUC4-fitted models with more predictors and slightly better-accuracy inputs decrease in performance when migrated to GRDC, although most of them are still better than the original Budyko (Figure 4). In the 14 models inspected (the 0th is Budyko itself), models #1–6 have 1–6 predictors, respectively, all of which include *A/P*. Model #1 with *A/P* as the only predictor contributes the most significant error reduction, while additional predictors have limited impact. When migrated to GRDC ($GRDC^M$), model #1 still reduces the error significantly, and this reduction is similar to that achieved by directly fitting the model to the GRDC data set ($GRDC^F$). With each additional predictor added (models #2–6), RMSE decreases slightly for HUC4 and somewhat more notably for $GRDC^F$. However, RMSE increases with the same models for $GRDC^M$. When *A/P* is absent (models #10–13), RMSE increases significantly for both HUC4 and GRDC. RMSE($GRDC^M$) even exceeds the Budyko model itself. From Table 1, the coefficient for ξ in GRDC-fitted model #10 is 0, suggesting seasonality plays different roles in different regions so that no consistent pattern exists on a global scale. In addition, for many models in Table 1, the coefficients for some other factors, e.g., γ and *S/P*, could switch signs between HUC4 and GRDC. This means their influences on *E* are region-specific. Model #13 is most interesting: with all five predictors other than *A/P*, this model is able to bring RMSE(HUC4) down to a level similar to model #1, but when migrated to GRDC, the error is even larger than original Budyko. We conclude that the coefficients estimated for these physical parameters are not portable across regions, consistent with previous research [*Xu et al*., 2013]. In contrast, *A/P* is an effective and crucial predictor whose coefficient is transferrable between regions.

When migrating HUC4-fitted models to the USGS basins, the performance gradually degrades for smaller basins (Figure 5). At the largest scale (20×10^3 km^2), R^2 is 0.85 for model #1 (with *A/P* only) while it is just 0.69 for Budyko itself. At around 4×10^2 km^2, R^2 decreases rapidly and for the original Budyko, it drops below 0.3. In terms of RMSE for *E/P*, the gap between model #1 and Budyko curve remains similar across scales, suggesting that the deterioration is not primarily due to overfitting or using GRACE signal from the enclosing HUC4. Such decrease may be due to the scale limitation of the Budyko hypothesis itself, but may also be partly attributed to the decrease in quality of climate input data. Our input data were obtained from NLDAS with a 1/8° spatial resolution (approximately 100 km^2 at 40°N and 144 km^2 at 30°N), which corresponds to the point when RMSE drops below 0.3.

Adding predictors to *A/P* does not noticeably improve performance for either GRDC or USGS. The overfitting problem is mild with models containing *A/P*, but significant for those without it. We notice that the model

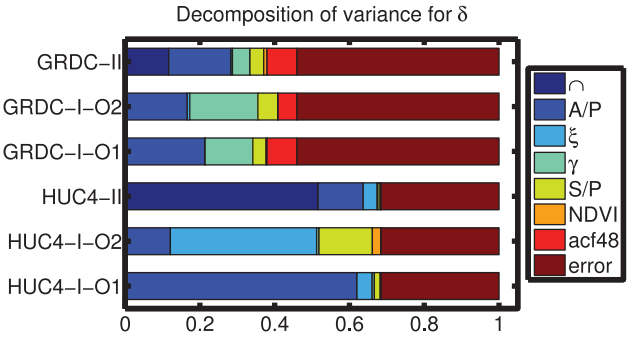

Figure 3. Variance decomposition to surrogate indicators and physical factors using ANCOVA for HUC4 and GRDC data sets. The fraction of the horizontal bars occupied by a factor indicates the variance explained by this factor. Type I Sum of Squares (SS) is "sequential" so the order of factors influences the results, unlike type II, which excludes main effects of other factors. The order in O1 is A/P, ξ, γ, S/P, NDVI, and acf_{48}, while O2 is ξ, γ, S/P, NDVI, acf_{48}, and A/P. The joint symbol, $\cap$, stands for the part of variance with attributed to more than one factors. As ξ and S/P occupy small fractions in O1 but more noticeable in O2, we conclude that A/P encompasses ξ and S/P, while the latter two constitute a large fraction, although not all of the predictive power of A/P.

with $A/P+\gamma+S/P+acf48$ remains the best model across all USGS basin scales, while when migrated to GRDC the overfitting is noticeable. We offer three explanations for the overfitting when migrating to GRDC: (1) catchments in the world experience more diverse climatic, geologic and human modification conditions than the USA, so the coefficients learned from HUC4 for the attributes other than *A/P* are not general enough. This is most apparent from the seasonality index, which has a coefficient of nearly 0 in the GRDC data set; (2) different data sets (NLDAS versus GLDAS + TRMM) have different biases in various regions; (3) catchment coevolution invalidates a linear correction model based on the physical factors.

The coevolution theory may help explain contrasts between the overfitting of physical factors and the portability of *A/P*-based models. The abiotic and biotic systems (e.g., vegetation, soil, topography, and landforms) coevolve and adapt to each other and climate conditions such as P-E_p phase differences [*Gentine et al.*, 2012; *Wagener et al.*, 2013]. *Troch et al.* [2013] argued that catchment characteristics coevolve with climate to produce the manifested E pattern such as the Budyko curve. Assuming this theory has validity, when climate changes, the system responds with a variety of intertwined and initial-value- and path-dependent changes in different physical factors. We have limited capability in capturing changes in all these factors. It is challenging to observe or too numerous to analyze robustly, when each factor only accounts for a small part of the variability. For example, climate change may cause a change in ξ, and then coevolution induces changes in other factors such as soil and vegetation (leaf and root). While changes in leaf status can be measured by *NDVI* and captured in our linear model, soil and root are much more difficult to measure. Linear correction models may fail to generalize due to missing factors. However, *A/P* as a surrogate captures some of the effects of all these responses because *A/P* covaries with the departure. Therefore, the existence of partial coevolution might be the reason why the surrogate model is favored over linear models with physical factors. As a side note, the coevolution theory above leads to the surmise that a departure from Budyko indicates incomplete coevolution, that is, if given sufficiently long time, these basins will eventually return to the Budyko curve.

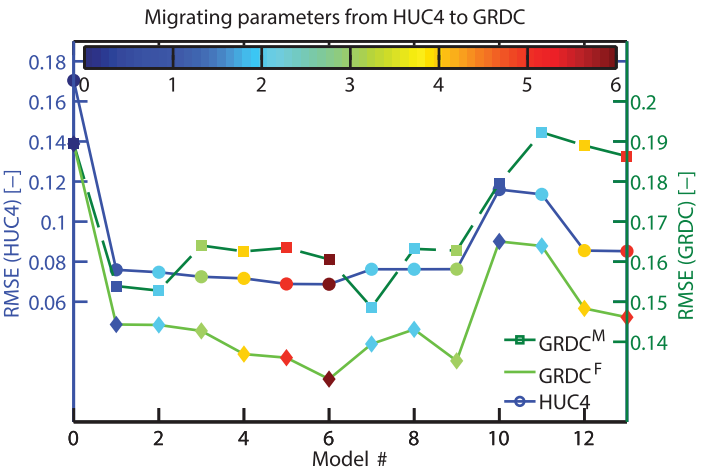

Figure 4. Errors of the models #0 through #13 (Table 1) when they are fitted to the HUC4 data (blue line) and their coefficients are migrated to the GRDC data set ($GRDC^M$ green solid line), compared to when the models that are directly fitted to the GRDC data ($GRDC^F$, green dashed line). Colors of the markers indicate the number of predictors. Note HUC4 and GRDC RMSEs are on two different y axes. We note that as indicators are added in models 1–6, they slightly reduce errors in HUC4 but increase errors when parameters are migrated to GRDC. In addition, while the model with A/P is transferrable from HUC4 to GRDC, models without A/P performs poorly when migrated. The evaluation period is October 2002 to September 2010, because GRDC is only available before September 2010.

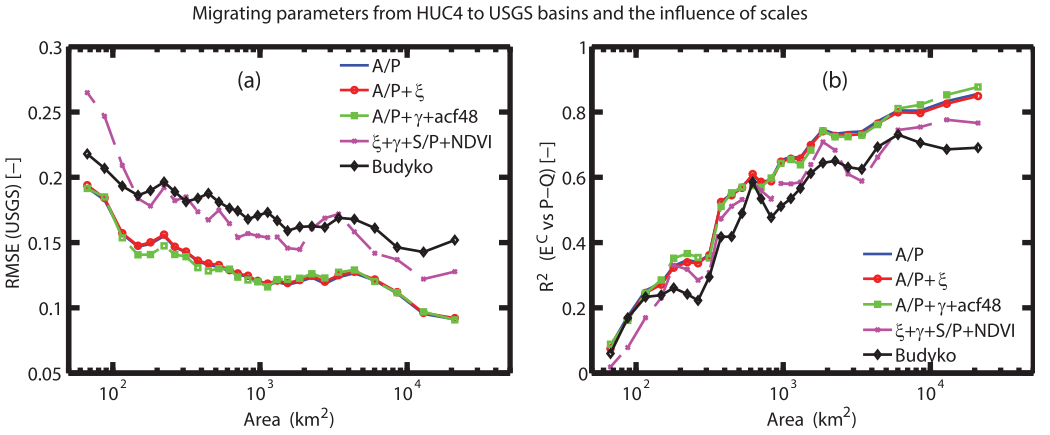

Figure 5. The scale-dependent performance of five models when HUC4-fitted models are migrated to the USGS GAGES-II basins. (a) RMSE of evaporation ratio (blue line is covered by red); (b) R2 between observed versus predicted E. The correction formula with A/P are transferrable to >1000 km2 basins while the one without A/P performed only slightly better than Budyko itself. The gradual degradation in performance toward smaller scales with the GRACE-assisted models is similar to that of the Budyko itself, indicating the degradation is due to either data limitation or inherent loss of accuracy of the Budyko equation for smaller basins, rather than due to the correction formula.

3.4. Comparison With Other *E* Products

We use model #9 calibrated based on GRDC data, i.e., with *A/P*, interannual variability (γ) and fraction of *P* as snow (acf_{48}) as predictors, to produce a world long-term annual *E*, termed E^C (Figure 6a). Overall, E^C had slighter higher correlation coefficient than E^{PJ}, while both perform much better than the original Budyko estimates (Figure 7a). Toward the higher *E* ranges (in the Amazon basin), E^{PJ} tends to overestimate *E* while E^C tends to underestimate *E*. The gridded product was compared with E^{PJ} and E^{GLDAS} (Figures 6b, 6c, 7b, and 7c). We note that in many places in the world the differences are small, and the difference between E^C and the two data products are smaller than that between themselves (Figures 7b and 7c). Regions with noticeable differences include the Amazon forest, Southern Brazil, central Africa, Southeast Asia, Northern Australia and Japan. In some of these regions (central Amazon, central and central-south Africa), three data sets all differ, but E^C is between E^{GLDAS} and E^{PJ}, and it agrees with one more than the other. In some regions (Southwest Amazon, Southeast Asia, Japan, and Northern Australia), E^C is likely to be in error due to GRACE error contamination. We discuss these differences in the following.

In most parts of the Amazon, E^C is substantially (0–400 mm yr^{-1}) smaller than E^{PJ} but somewhat larger (0–250 mm yr^{-1}) than E^{GLDAS}, except in the southwest. Judging from the limited GRDC comparisons (Figure 7a toward to the high *E* range), true *E* in this region is likely in the middle between E^C and E^{PJ}, and E^{GLDAS} is likely the most biased potentially because of underestimating evaporation of canopy interception. Both E^C and E^{GLDAS} can suffer from errors in precipitation which can be significant in this region [*Zhou et al.*, 2012]. In addition, as shown previously, using *NDVI* as physical predictors did not help with this issue. It is possible that the satellite-based sensing of *NDVI* and LAI are saturated in the Amazon, invalidating a linear correction formula. In the southwest Amazon, E^C is slightly less than both and is likely to be in error here. As we can see from the GRACE error map (Figure A3a), the southwest Amazon has large leakage errors. In fact, it is clear from Figures 7b and 7c that in regions with large combined GRACE errors, E^C is likely to be biased.

In central and central-south Africa (immediately beneath the Sahara), E^C is slightly less than E^{PJ} yet it is noticeably larger than E^{GLDAS}. The agreement between E^C and E^{PJ}, the small GRACE error in this region and the large difference from E^{GLDAS} suggest that E^{GLDAS} is in error here. In contrast, in Southeast Asia and Northern Australia, E^C is bigger than both E^{PJ} and E^{GLDAS}. Even after the interpolation attempt, the error is still large in this region. Given the large leakage error in this region, E^C most likely has a large positive bias here. E^{PJ} data are missing for Japan, but E^C is also unreliable here due to large measurement error.

4. Limitations and Future Work

As discussed above, in regions where GRACE errors are large (e.g., Southeast Asia and Japan), the correction formula might be contaminated by errors. In the future, it should be possible to merge the GRACE-assisted

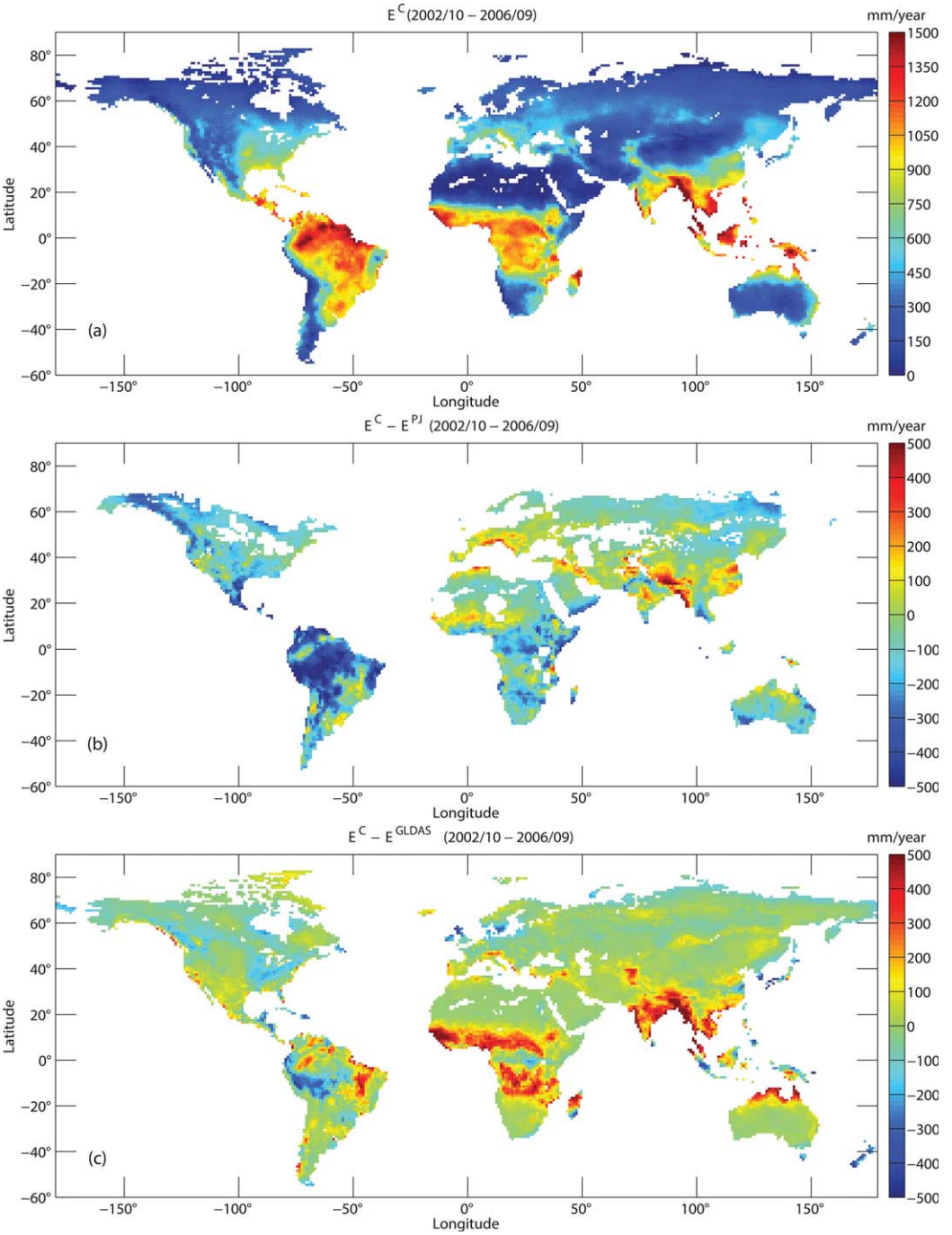

Figure 6. (a) World annual average GRACE-corrected evapotranspiration, E^C, for 2002–2006 based on Budyko and linear model with A/P, interannual variability index (γ) and S/P, after GRACE error limiting; (b) difference between E^C and E^{PJ} (white blanks are due to missing data in E^{PJ}); and (c) difference between E^C and E^{GLDAS}.

product with others using data fusion techniques that weigh GRACE-corrected data using combined GRACE errors. With advances in algorithms, we might be able to obtain improved GRACE estimates with lower errors, or new algorithms to reduce the area of regions with large errors. Because all inputs are available on a monthly time scale, it should be possible to produce annual *E* data set, instead of a long-term average. However, extracting meaningful amplitude for each year requires additional work, which we leave for the next stage.

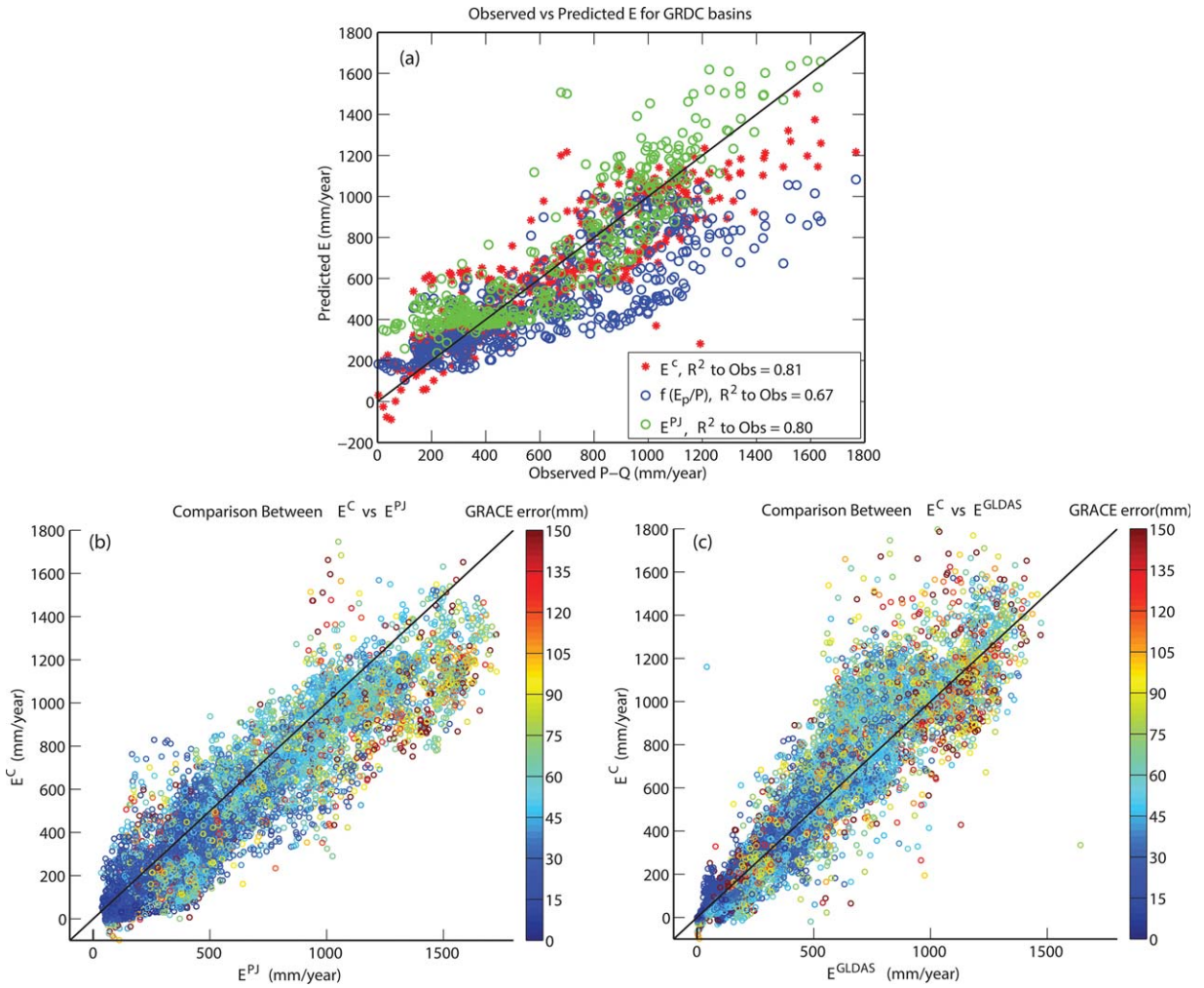

Figure 7. (a) Comparing GRACE-corrected product (E^C) with E^{PJ} and Budyko itself in GRDC basins for October 2002 to September 2006 (E^{PJ} is available only before September 2006). Budyko has negative biases while E^{PJ} has positive biases in high E regions. (b–c) Comparison of E^C with E^{PJ} and GLDAS evapotranspiration. The root-mean-squared-difference is 175 mm/yr between E^C and E^{PJ}, 131 mm/yr between E^C and E^{GLDAS}, and 189 mm/yr between E^{PJ} and E^{GLDAS}. We note that red-colored points (large GRACE errors) tend to be the most distant to the 1-to-1 line, indicating in these regions E^C is unreliable due to GRACE error contamination.

5. Conclusions

Given the large number of *E* products available for the world and leakage/measurement errors facing GRACE, our estimate is more useful in a data-fusion setting than being used as a standalone product. However, the novelty of this work resides with the advances in understanding the relationship between storage amplitude and *E*, using Budyko as a reference condition, the mechanisms influencing TWSA amplitude, the hydrologic signatures from GRACE and the transferability of the models embodying different hypotheses. To our knowledge, the present work is the first one examining the physical significance of TWSA amplitude as a fraction of precipitation (*A/P*) and its controls by seasonality and snow, and the first one to link it to the departure from the Budyko curve. The surrogate indicator *A/P* is a powerful and portable predictor for the departure from the Budyko curve. Migrating models from HUC4 to GRDC and USGS data sets, the models with *A/P* are more transferrable than those without it. We argue the overfitting with physical-factor-based linear models may be due to partial coevolution of basin characteristics with climate, which can be difficult to fully capture. Although GRACE has a coarse spatial resolution, the methodology works well for basins above 1000 km^2 in the USA, which is much smaller than the GRACE footprint. The gradual degradation in performance toward smaller scales is not due to the GRACE-based correction formula, but Budyko itself and data limitations. Compared to two different global *E* products, in many regions in the world, our improved estimate (E^C) is either similar to both,

between the two, or agree with one more than the other. The errors with E^C is related to GRACE measurement and leakage errors. Southeast Asia, Southwest Amazon, Northern Australia, and Japan are regions where E^C most likely has large biases.

Appendix A: Different Interpretations of the Linear Regression Equation

When $\boldsymbol{x}$ is composed of only independent physical factors, equation (7) can be interpreted as having made the following hypothesis: the Budyko formula describes the E of a standard reference basin, $\overline{\boldsymbol{x}}$, ($\overline{\boldsymbol{x}}$

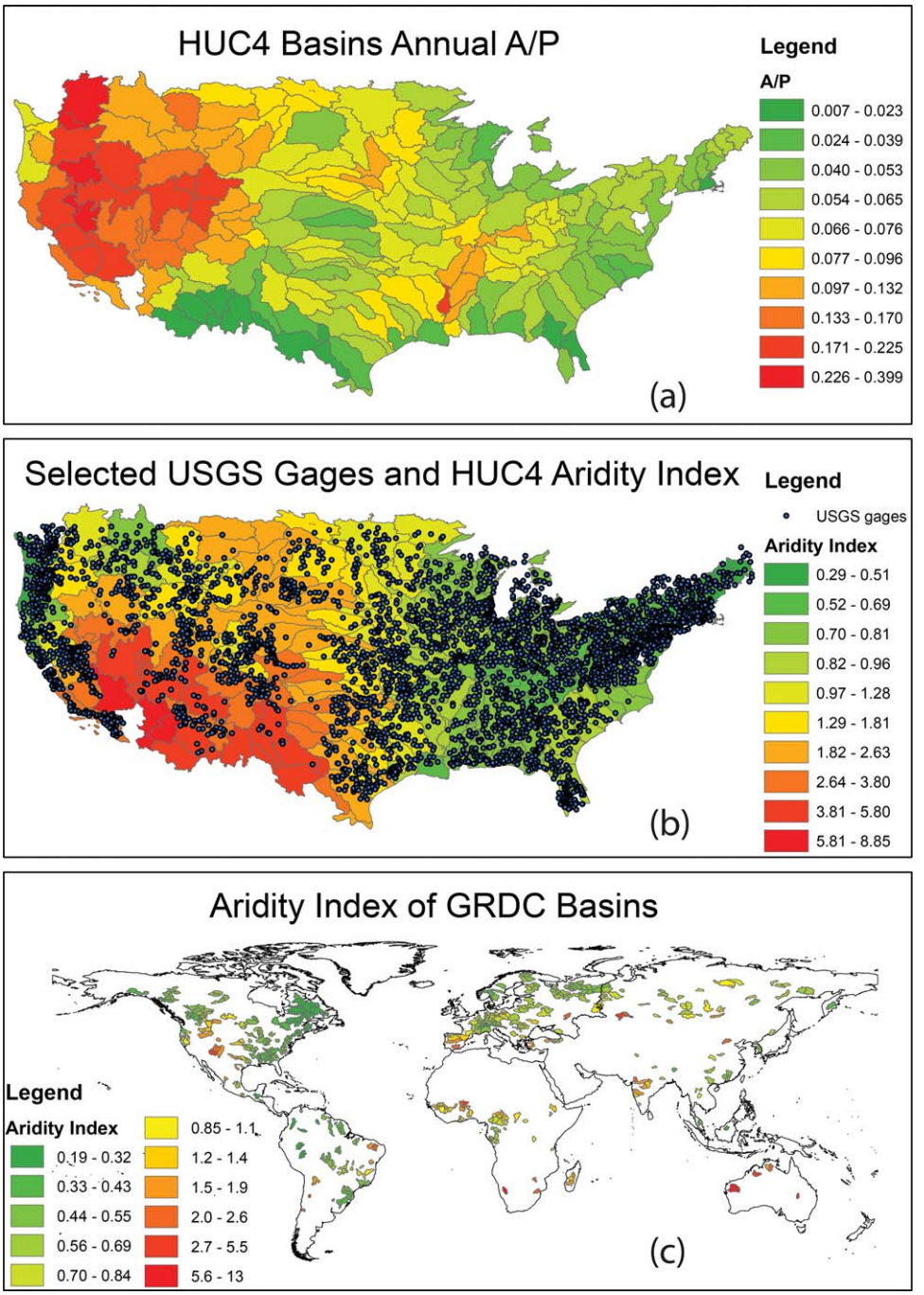

Figure B1. (a) A/P for HUC4; (b) locations of USGS gages used in the study; (c) maps of GRDC basins [*Global Runoff Data Center*, 2011] used in the study, with colors indicating the aridity index.

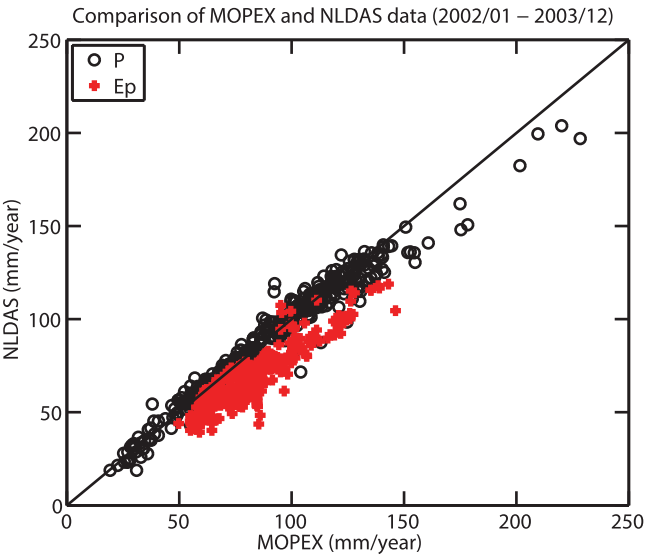

Figure B2. Comparisons of annual average fluxes between HUC4 and MOPEX basins for the period January 2002 to December 2013.

stands for the reference values of a comprehensive set of physical factors, e.g., vegetation cover, terrain slope, phase shift between *P* and *Ep*, etc.), surrounding which the actual *E* changes smoothly as a function of the physical factors. The reference basin represents average conditions, which may vary as a function of the aridity index, of world catchments. The smoothness assumption allows us to approximate the deviation from the reference state using a linear formula of the change in the factors, i.e., employing Taylor Series expansion

$$\delta(\boldsymbol{x})=\frac{E}{P}-f\left(\frac{E_p}{P}\right)=\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{x}-\overline{\boldsymbol{x}})+\varepsilon. \quad \text{(A1)}$$

When $\overline{\boldsymbol{x}}$ are constants, they need not be estimated independently, but can be lumped into the one constant in the linear regression, i.e., in terms of data fitting, equation (A1) is equivalent to equation (7).

When $\boldsymbol{x}$ also contains surrogate indices, e.g., *A/P* and γ, it has a different physical meaning. If, in addition to influencing *E/P*, $\boldsymbol{x}$ also influence the surrogate indices, say *A/P*, we can write an equation similar to equation (A1)

$$\frac{\Delta A}{P}=\frac{A}{P}-\frac{A_0}{P}=\boldsymbol{b}^{\boldsymbol{T}}(\boldsymbol{x}-\overline{\boldsymbol{x}})+\varepsilon, \quad \text{(A2)}$$

where A_0 is a reference amplitude, at which the amplitude-based correction is 0. We can split $\overline{\boldsymbol{x}}$ into two components, $\overline{\boldsymbol{x}}=\overline{\boldsymbol{x}_0}+\overline{\boldsymbol{x}_C}$, so the above equations can be rewritten as

$$\delta=\boldsymbol{a}_0^{\boldsymbol{T}}(\boldsymbol{x}_0-\overline{\boldsymbol{x}_0})+\boldsymbol{a}_{\boldsymbol{C}}^{\boldsymbol{T}}(\boldsymbol{x}_{\boldsymbol{C}}-\overline{\boldsymbol{x}_{\boldsymbol{C}}})+\varepsilon, \quad \text{(A3a)}$$

$$\frac{A}{P}-\frac{A_0}{P}=\boldsymbol{b}_0^{\boldsymbol{T}}(\boldsymbol{x}_0-\overline{\boldsymbol{x}_0})+\boldsymbol{b}_{\boldsymbol{C}}^{\boldsymbol{T}}(\boldsymbol{x}_{\boldsymbol{C}}-\overline{\boldsymbol{x}_{\boldsymbol{C}}})+\varepsilon, \quad \text{(A3b)}$$

where $\boldsymbol{x}_{\boldsymbol{C}}$ are major climate or basin characteristics that can be conveniently computed with available data for ungauged basins, e.g., phase shift between *P* and E_p or fraction of precipitation as snow, fraction of precipitation falling as snow (*S/P*) or vegetation indices (*NDVI*). We have tested using the $\overline{NDVI}$ as an aridity-dependent variable, i.e., fitting the mean *NDVI* to aridity value, but this did not improve our results. $\boldsymbol{x}_0$ is central to our method: these are a set of factors that influence both δ and $\Delta A/P$ and we can find an approximate, effective ratio β between $\boldsymbol{b}_0$ and $\boldsymbol{a}_0$, i.e., $\beta=\boldsymbol{a}_0/\boldsymbol{b}_0$. As a result, $\Delta A/P$ can be used as a surrogate for $\boldsymbol{x}_0$, and we can rewrite equation (A3a) as

$$\delta^*(\mathrm{x})=\beta\frac{\Delta A}{P}+(\boldsymbol{a}_{\boldsymbol{C}}^{\boldsymbol{T}}-\beta\boldsymbol{b}_{\boldsymbol{C}}^{\boldsymbol{T}})(\boldsymbol{x}_{\boldsymbol{C}}-\overline{\boldsymbol{x}_{\boldsymbol{C}}}). \quad \text{(A4)}$$

This formula gives some flexibility in assigning a factor into either $\boldsymbol{x}_0$ or $\boldsymbol{x}_{\boldsymbol{C}}$. Although a fixed ratio $\beta=\boldsymbol{a}_0/\boldsymbol{b}_0$ seems a strong assumption, in reality, many factors exert weak controls, or they have strong controls but with limited variability, and we can lump them into an effective parameter to be represented by $\Delta A/P$. At the extreme, we can merge all of $\boldsymbol{x}_{\boldsymbol{E}}$ and $\boldsymbol{x}_{\boldsymbol{C}}$ into $\boldsymbol{x}_0$. Equation (A4) becomes $\delta\left(\frac{E_p}{P},\ x\right)=\beta\frac{\Delta A}{P}+\varepsilon$. When we test equation (A4), the prediction of δ turns into a linear regression problem between $\frac{\Delta A}{P}$ and δ so again it is equivalent to equation (7) for parameter estimation purpose.

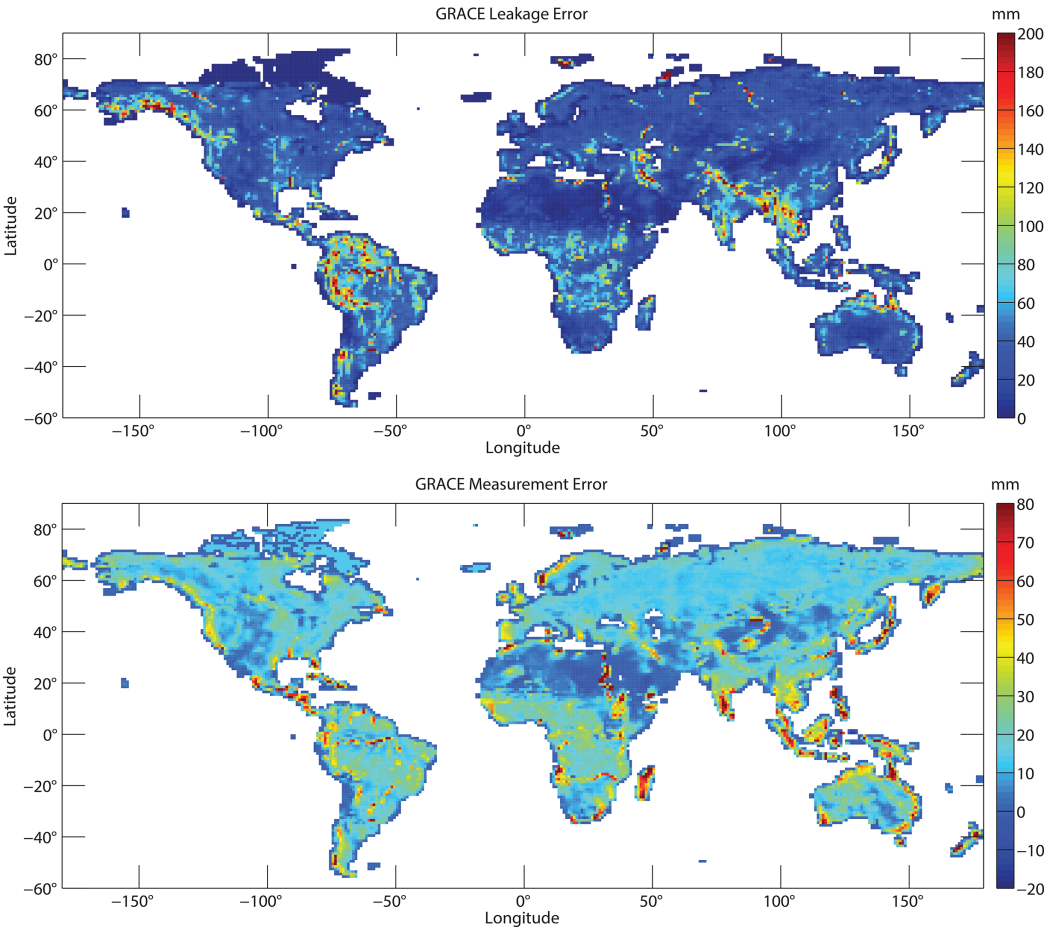

Figure B3. GRACE leakage (top) and measurement (bottom) errors.

Appendix B: Supporting Figures for Variables Discussed

The boundaries of HUC4 and GRDC datasets, long-term average aridity index and A/P for HUC4 are presented in Figure B1. To bridge the communities that use MOPEX and NLDAS datasets, we show their comparisons in Figure B2. Finally, Figure B3 presents GRACE leakage and measurement errors. From this figure, we notice that Northwestern coast of North America, Andes, South Asia, Japan, Indonesia, Madagascar and Northern Australia are all regions with relatively large GRACE errors.

Acknowledgments

This work was supported by Office of Biological and Environmental Research of the US Department of Energy under contract DE-SC0010620. We thank David Wolock from USGS for providing shapefiles for the USGS basins. Data generated from this study are presented in figure format in the paper, and the data sets can be requested from the corresponding author. We thank Murugesu Sivapalan for some useful discussion about incomplete coevolution. J.B.F. contributed to this work from the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Constructive comments from anonymous reviewers and the Associate Editor have helped to improve the manuscript.

References

Arora, V. K. (2002), The use of the aridity index to assess climate change effect on annual runoff, *J. Hydrol.*, *265*(1–4), 164–177, doi:10.1016/S0022-1694(02)00101-4.

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

Berghuijs, W. R., M. Sivapalan, R. A. Woods, and H. H. G. Savenije (2014b), Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales, *Water Resour. Res.*, *50*, 5638–5661, doi:10.1002/2014WR015692.

Budyko, M. I. (1948), *Evaporation Under Natural Conditions*, Gidrometeorizdat, Jerusalem.

Buermann, W. (2002), Analysis of a multiyear global vegetation leaf area index data set, *J. Geophys. Res.*, *107*(D22), 4646, doi:10.1029/2001JD000975.

Carmona, A. M., M. Sivapalan, M. A. Yaeger, and G. Poveda (2014), Regional patterns of interannual variability of catchment water balances across the continental U.S.: A Budyko framework, *Water Resour. Res.*, *50*, 9177–9193, doi:10.1002/2014WR016013.

Chen, X., N. Alimohammadi, and D. Wang (2013), Modeling interannual variability of seasonal evaporation and storage change based on the extended Budyko framework, *Water Resour. Res.*, *49*, 6067–6078, doi:10.1002/wrcr.20493.

Chen, Y., et al. (2014), Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in China, *Remote Sens. Environ.*, *140*, 279–293, doi:10.1016/j.rse.2013.08.045.

Clark, M. P., et al. (2015), Improving the representation of hydrologic processes in Earth System Models, *Water Resour. Res.*, *51*, 5929–5956, doi:10.1002/2015WR017096.

Döll, P., H. Müller Schmied, C. Schuh, F. T. Portmann, and A. Eicker (2014), Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites, *Water Resour. Res.*, *50*, 5698–5720, doi:10.1002/2014WR015595.

Duan, Q., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*(1–2), 3–17, doi:10.1016/j.jhydrol.2005.07.031.
Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley (2003), Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, *J. Geophys. Res.*, *108*(D22), 8851, doi:10.1029/2002JD003296.
Emery, W. J., and R. E. Thomson (2004), *Data Analysis Methods in Physical Oceanography*, Elsevier, Amsterdam, Netherlands.
Falcone, J. (2011), *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow*, U.S. Geol. Surv., Reston, Va. [Available at http://water.usgs.gov/lookup/getspatial?gagesII_Sept2011.]
Famiglietti, J. S., M. Lo, S. L. Ho, J. Bethune, K. J. Anderson, T. H. Syed, S. C. Swenson, C. R. de Linage, and M. Rodell (2011), Satellites measure recent rates of groundwater depletion in California's Central Valley, *Geophys. Res. Lett.*, *38*, L03403, doi:10.1029/2010GL046442.
Fatichi, S., et al. (2016), An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, *J. Hydrol.*, *537*, 45–60, doi:10.1016/j.jhydrol.2016.03.026.
Fisher, J. B., K. P. Tu, and D. D. Baldocchi (2008), Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites, *Remote Sens. Environ.*, *112*(3), 901–919, doi:10.1016/j.rse.2007.06.025.
Fox, J. (2008), *Applied Regression Analysis and Generalized Linear Models*, 2nd ed., SAGE, Los Angeles, Calif.
Fu, B. P. (1981), On the calculation of the evaporation from land surface [in Chinese], *Sci. Atmos. Sin.*, *5*(1), 23–31.
Gentine, P., P. D'Odorico, B. R. Lintner, G. Sivandran, and G. Salvucci (2012), Interdependence of climate, soil, and vegetation as constrained by the Budyko curve, *Geophys. Res. Lett.*, *39*, L19404, doi:10.1029/2012GL053492.
Gerrits, A. M. J., H. H. G. Savenije, E. J. M. Veling, and L. Pfister (2009), Analytical derivation of the Budyko curve based on rainfall characteristics and a simple evaporation model, *Water Resour. Res.*, *45*, W04403, doi:10.1029/2008WR007308.
Global Runoff Data Center (2011), Watershed boundaries of GRDC Stations, GRDC in the Bundesanstalt fuer Gewaesserkunde, 56068 Kolenz, Germany. [Available at http://grdc.bafg.de.].
Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.
Hrachowitz, M., et al. (2013), A decade of Predictions in Ungauged Basins (PUB)—A review, *Hydrol. Sci. J.*, *58*(6), 1198–1255, doi:10.1080/02626667.2013.803183.
Huang, Z., Y. Pan, H. Gong, P. J.-F. Yeh, X. Li, D. Zhou, and W. Zhao (2015), Sub-regional scale groundwater depletion detected by GRACE for both shallow and deep aquifers in North China Plain, *Geophys. Res. Lett.*, *42*, 1791–1799, doi:10.1002/2014GL062498.
Huffman, G. J., R. F. Adler, P. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolf, and U. Schneider (1997), The Global Precipitation Climatology Project (GPCP) Combined Precipitation Dataset, *Bull. Am. Meteorol. Soc.*, *78*(1), 5–20, doi:10.1175/1520-0477(1997)078<0005:TGPCPG>2.0.CO;2.
Istanbulluoglu, E., T. Wang, O. M. Wright, and J. D. Lenters (2012), Interpretation of hydrologic trends from a water balance perspective: The role of groundwater storage in the Budyko hypothesis, *Water Resour. Res.*, *48*, W00H16, doi:10.1029/2010WR010100.
Landerer, F. W., and S. C. Swenson (2012), Accuracy of scaled GRACE terrestrial water storage estimates, *Water Resour. Res.*, *48*, W04531, doi:10.1029/2011WR011453.
Keppel, G., and T. D. Wickens (2004), *Design and Analysis: A Researcher's Handbook*, 4th ed., Pearson Prentice Hall, Upper Saddle River, N. J.
Li, D. (2014), Assessing the impact of interannual variability of precipitation and potential evaporation on evapotranspiration, *Adv. Water Resour.*, *70*, 1–11, doi:10.1016/j.advwatres.2014.04.012.
Li, D., M. Pan, Z. Cong, L. Zhang, and E. Wood (2013), Vegetation control on water and energy balance within the Budyko framework, *Water Resour. Res.*, *49*, 969–976, doi:10.1002/wrcr.20107.
Li, H.-Y., M. Sivapalan, F. Tian, and C. Harman (2014), Functional approach to exploring climatic and landscape controls of runoff generation: 1. Behavioral constraints on runoff volume, *Water Resour. Res.*, *50*, 9300–9322, doi:10.1002/2014WR016307.
Lo, M.-H., J. S. Famiglietti, P. J.-F. Yeh, and T. H. Syed (2010), Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data, *Water Resour. Res.*, *46*, W05517, doi:10.1029/2009WR007855.
Long, D., Y. Shen, A. Sun, Y. Hong, L. Longuevergne, Y. Yang, B. Li, and L. Chen (2014a), Drought and flood monitoring for a large karst plateau in Southwest China using extended GRACE data, *Remote Sens. Environ.*, *155*, 145–160, doi:10.1016/j.rse.2014.08.006.
Long, D., L. Longuevergne, and B. R. Scanlon (2014b), Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites, *Water Resour. Res.*, *50*, 1131–1151, doi:10.1002/2013WR014581.
McCabe, M. F., A. Ershadi, C. Jimenez, D. G. Miralles, D. Michel, and E. F. Wood (2015), The GEWEX LandFlux project: Evaluation of model evaporation using tower-based and globally-gridded forcing data, *Geosci. Model Dev. Discuss.*, *8*(8), 6809–6866, doi:10.5194/gmdd-8-6809-2015.
Milly, P. C. D. (1994), Climate, soil water storage, and the average annual water balance, *Water Resour. Res.*, *30*(7), 2143–2156, doi:10.1029/94WR00586.
Miralles, D. G., et al. (2015), The WACMOS-ET project – Part 2: Evaluation of global terrestrial evaporation data sets, *Hydrol. Earth Syst. Sci. Discuss.*, *12*(10), 10651–10700, doi:10.5194/hessd-12-10651-2015.
Mu, Q., M. Zhao, and S. W. Running (2011), Improvements to a MODIS global terrestrial evapotranspiration algorithm, *Remote Sens. Environ.*, *115*(8), 1781–1800, doi:10.1016/j.rse.2011.02.019.
Niu, J., C. Shen, S.-G. Li, and M. S. Phanikumar (2014), Quantifying storage changes in regional Great Lakes watersheds using a coupled subsurface-land surface process model and GRACE, MODIS products, *Water Resour. Res.*, *50*, 7359–7377, doi:10.1002/2014WR015589.
Oleson, K. W., et al. (2013), Technical description of version 4.5 of the Community Land Model (CLM), *NCAR/TN-503+STR, NCAR Tech. Note*, UCAR/NCAR, Boulder, Colo. [Available at http://www.cesm.ucar.edu/models/cesm1.2/clm/CLM45_Tech_Note.pdf.]
Pike, J. G. (1964), The estimation of annual run-off from meteorological data in a tropical climate, *J. Hydrol.*, *2*(2), 116–123, doi:10.1016/0022-1694(64)90022-8.
Potter, N. J., L. Zhang, P. C. D. Milly, T. A. McMahon, and A. J. Jakeman (2005), Effects of rainfall seasonality and soil moisture capacity on mean annual water balance for Australian catchments, *Water Resour. Res.*, *41*, W06007, doi:10.1029/2004WR003697.
Reager, J. T., and J. S. Famiglietti (2013), Characteristic mega-basin water storage behavior using GRACE, *Water Resour. Res.*, *49*, 3314–3329, doi:10.1002/wrcr.20264.
Reager, J. T., B. F. Thomas, and J. S. Famiglietti (2014), River basin flood potential inferred using GRACE gravity observations at several months lead time, *Nat. Geosci.*, *7*(8), 588–592, doi:10.1038/ngeo2203.
Rodell, M., et al. (2004), The Global Land Data Assimilation System, *Bull. Am. Meteorol. Soc.*, *85*(3), 381–394, doi:10.1175/BAMS-85-3-381.
Sawicz, K. A., C. Kelleher, T. Wagener, P. Troch, M. Sivapalan, and G. Carrillo (2014), Characterizing hydrologic change through catchment classification, *Hydrol. Earth Syst. Sci.*, *18*(1), 273–285, doi:10.5194/hess-18-273-2014.

Scanlon, B. R., C. C. Faunt, L. Longuevergne, R. C. Reedy, W. M. Alley, V. L. McGuire, and P. B. McMahon (2012), Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley, *Proc. Natl. Acad. Sci. USA*, *109*(24), 9320–9325, doi:10.1073/pnas.1200311109.

Shen, C., J. Niu, and M. Phanikumar (2013), Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface-land surface processes model, *Water Resour. Res.*, *49*, 2552–2572, doi:10.1002/wrcr.20189.

Shen, C., J. Niu, and K. Fang (2014), Quantifying the effects of data integration algorithms on the outcomes of a subsurface-land surface processes model, *Environ. Model. Software*, *59*, 146–161, doi:10.1016/j.envsoft.2014.05.006.

Shen, C., W. J. Riley, K. M. Smithgall, J. M. Melack, and K. Fang (2016), The fan of influence of streams and channel feedbacks to simulated land surface water and carbon dynamics, *Water Resour. Res.*, *52*, 880–902, doi:10.1002/2015WR018086.

Shuttleworth (1993), Evaporation, in *Handbook of Hydrology*, edited by D. Maidment, pp. 4.1–4.53, McGraw-Hill, N. Y.

Sivapalan, M., M. A. Yaeger, C. J. Harman, X. Xu, and P. A. Troch (2011), Functional model of water balance variability at the catchment scale: 1. Evidence of hydrologic similarity and space-time symmetry, *Water Resour. Res.*, *47*, W02522, doi:10.1029/2010WR009568.

Slater, L. J., and M. B. Singer (2013), Imprint of climate and climate change in alluvial riverbeds: Continental United States, 1950-2011, *Geology*, *41*(5), 595–598, doi:10.1130/G34070.1.

Swenson, S. (2012), *GRACE monthly land water mass grids ASCII release 5.0*, Jet Propul. Lab., Pasadena, Calif. [Available at http://dx.doi.org/10.5067/TELND-TX005.]

Swenson, S. (2014), *GRACE Monthly Mass Grids - Land*, Jet Propul. Lab., Pasadena, Calif. [Available at http://grace.jpl.nasa.gov.].

Swenson, S., and J. Wahr (2006), Post-processing removal of correlated errors in GRACE data, *Geophys. Res. Lett.*, *33*, L08402, doi:10.1029/2005GL025285.

Tapley, B. D., S. Bettadpur, M. Watkins, and C. Reigber (2004), The gravity recovery and climate experiment: Mission overview and early results, *Geophys. Res. Lett.*, *31*, L09607, doi:10.1029/2004GL019920.

Troch, P. A., G. Carrillo, M. Sivapalan, T. Wagener, and K. Sawicz (2013), Climate-vegetation-soil interactions and long-term hydrologic partitioning: Signatures of catchment co-evolution, *Hydrol. Earth Syst. Sci.*, *17*(6), 2209–2217, doi:10.5194/hess-17-2209-2013.

Vinukollu, R. K., E. F. Wood, C. R. Ferguson, and J. B. Fisher (2011), Global estimates of evapotranspiration for climate studies using multi-sensor remote sensing data: Evaluation of three process-based approaches, *Remote Sens. Environ.*, *115*(3), 801–823, doi:10.1016/j.rse.2010.11.006.

Vogel, R. M., and A. Sankarasubramanian (2003), Validation of a watershed model without calibration, *Water Resour. Res.*, *39*(10), 1292, doi:10.1029/2002WR001940.

Voss, K. A., J. S. Famiglietti, M. Lo, C. Linage, M. Rodell, and S. C. Swenson (2013), Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-Western Iran region, *Water Resour. Res.*, *49*, 904–914, doi:10.1002/wrcr.20078.

Wagener, T., G. Blöschl, D. C. Goodrich, H. V. Gupta, M. Sivapalan, Y. Tachikawa, P. A. Troch, and M. Weiler (2013), A synthesis framework for runoff prediction in ungauged basins, in *Runoff Prediction in Ungauged Basins. Synthesis Across Processes, Places and Scales*, edited by G. Bloschl et al., pp. 11–28, Cambridge Univ. Press, N. Y.

Wahr, J. (2004), Time-variable gravity from GRACE: First results, *Geophys. Res. Lett.*, *31*, L11501, doi:10.1029/2004GL019779.

Wahr, J., S. Swenson, and I. Velicogna (2006), Accuracy of GRACE mass estimates, *Geophys. Res. Lett.*, *33*, L06401, doi:10.1029/2005GL025305.

Wang, D., and Y. Tang (2014), A one-parameter Budyko model for water balance captures emergent behavior in Darwinian hydrologic models, *Geophys. Res. Lett.*, *41*, 4569–4577, doi:10.1002/2014GL060509.

Woods, R. A. (2009), Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks, *Adv. Water Resour.*, *32*(10), 1465–1481, doi:10.1016/j.advwatres.2009.06.011.

Xia, Y., 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, *J. Geophys. Res.*, *117*, D03109, doi:10.1029/2011JD016048.

Xu, X., W. Liu, B. R. Scanlon, L. Zhang, and M. Pan (2013), Local and global factors controlling water-energy balances within the Budyko framework, *Geophys. Res. Lett.*, *40*, 6123–6129, doi:10.1002/2013GL058324.

Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*(8), 1756–1774.

Yang, H., D. Yang, Z. Lei, and F. Sun (2008), New analytical derivation of the mean annual water-energy balance equation, *Water Resour. Res.*, *44*, W03410, doi:10.1029/2007WR006135.

Yang, H., D. Yang, and Q. Hu (2014), An error analysis of the Budyko hypothesis for assessing the contribution of climate change to runoff, *Water Resour. Res.*, *50*, 9620–9629, doi:10.1002/2014WR015451.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W09417, doi:10.1029/2007WR006716.

Zhang, L., W. R. Dawes, and G. R. Walker (2001), Response of mean annual evapotranspiration to vegetation changes at catchment scale, *Water Resour. Res.*, *37*(3), 701–708, doi:10.1029/2000WR900325.

Zhang, S., H. Yang, D. Yang, and A. W. Jayawardena (2015), Quantifying the effect of vegetation change on the regional water balance within the Budyko framework, *Geophys. Res. Lett.*, *43*, 1140–1148, doi:10.1002/2015GL066952.

Zhou, M. C., H. Ishidaira, H. P. Hapuarachchi, J. Magome, A. S. Kiem, and K. Takeuchi (2006), Estimating potential evapotranspiration using Shuttleworth–Wallace model and NOAA-AVHRR *NDVI* data to feed a distributed hydrological model over the Mekong River basin, *J. Hydrol.*, *327*(1–2), 151–173, doi:10.1016/j.jhydrol.2005.11.013.

Zhou, X., Y. Zhang, Y. Wang, H. Zhang, J. Vaze, L. Zhang, Y. Yang, and Y. Zhou (2012), Benchmarking global land surface models against the observed mean annual runoff from 150 large basins, *J. Hydrol.*, *470–471*, 269–279, doi:10.1016/j.jhydrol.2012.09.002.