AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR022723

Hydrologic Observation, Model, and Theory Congruence on Evapotranspiration Variance: Diagnosis of Multiple Observations and Land Surface Models

Ruijie Zeng[1,2] **and Ximing Cai**[2]

[1]Utah Water Research Laboratory, Department of Civil and Environmental Engineering, Utah State University, Logan, UT, USA, [2]Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Key Points:

- A diagnostic framework based on an observation-model-theory triplet is provided for cross-validating hydrologic theories, observations, and models
- The spatial patterns of ET variance assessed from four observation-based ET products are inconsistent; meanwhile, any of the four LSMs can be the *best* one for a certain set of reference observations
- Due to limited representation of human dimensions on ET and other hydrologic processes, LSMs are not able to accurately predict the ET variability in regions/aquifers heavily influenced by human activities, like Central Valley and High Plains aquifers

Supporting Information:

- Supporting Information S1

Correspondence to:
R. Zeng,
ruijie.zeng@usu.edu

Citation:
Zeng, R., & Cai, X. (2018). Hydrologic observation, model, and theory congruence on evapotranspiration variance: Diagnosis of multiple observations and land surface models. *Water Resources Research*, *54*. https://doi.org/10.1029/2018WR022723

Received 9 FEB 2018
Accepted 9 OCT 2018
Accepted article online 22 OCT 2018

Abstract This paper reconciles the state-of-the-art observations and simulations of evapotranspiration (*ET*) temporal variability through a diagnostic framework composed of an observation-model-theory triplet. Specifically, a confirmed theoretical tool, Evapotranspiration Temporal VARiance Decomposition (EVARD), is used as a benchmark to estimate ET monthly variance (σ^2_{ET}) across the contiguous United States (CONUS) with inputs including hydroclimatic observations, Gravity Recovery and Climate Experiment-based terrestrial water storage, four observation-based products ($\mathrm{ET_{RSUW}}$ by the University of Washington, $\mathrm{ET_{RSMOD16}}$ from MOD16 Global Terrestrial ET Data Set, $\mathrm{ET_{FLUXNET}}$ upscaled from of fluxtower observations, and $\mathrm{ET_{GLEAM}}$ from Global Land Evaporation Amsterdam Model), and four operational land surface models (LSMs: MOSAIC, NOAH, NOAH-MP, and VIC). Five experiments are systematically designed to evaluate and diagnose possible errors and uncertainties in ET temporal variance estimated by the four observation-based ET products and the four LSM simulations. Based on the results of these experiments, the following diagnostic hypotheses regarding the uncertainty of the observation-based ET products are illustrated: $\mathrm{ET_{RSUW}}$ captures the high σ^2_{ET} signals in the Midwest with negligible bias and moderate uncertainty over the contiguous United States; $\mathrm{ET_{FLUXNET}}$ systematically underestimates σ^2_{ET} over CONUS but with the lowest level of uncertainty; $\mathrm{ET_{RSMOD16}}$ has medium bias with the highest level of uncertainty, and the spatial distribution of high σ^2_{ET} signal from $\mathrm{ET_{RSMOD16}}$ is different from other estimates; $\mathrm{ET_{GLEAM}}$ has slight negative bias and medium uncertainty, and σ^2_{ET} in the West Coast is smaller than that from ETVARD. Regarding the LSMs, it is found that any of the four LSMs can be the *best* depending on a certain set of reference observations. The study reveals that LSMs have shown a reasonably worthy, though not perfect, capability in estimating ET and its variability in regions/aquifers with limited human interference. However, RS-based observations and theoretical estimates suggest that all the four LSMs examined in this study are not able to accurately predict the ET variability in regions/aquifers heavily influenced by human activities like Central Valley and High Plains aquifers; they all underestimate ET variability along the West Coast due to seasonal vegetation responses to Mediterranean climate and human water use. In addition, LSMs underestimate intraannual ET variance in California and the High Plains with underestimated terrestrial storage change components in ET variance, due to the inappropriate representation of groundwater pumping and its impact on ET and other hydrologic processes. This paper urges advancing hydrologic knowledge by finding congruence among models, data, and theories.

1. Introduction

Evapotranspiration (*ET*) is a key hydroecological process that couples water and energy budgets (Yang et al., 2008), links carbon and nutrient cycles (Porporato et al., 2015), and represents water consumption in food and biomass production (Housh et al., 2014). Numerous efforts have been made in hydrologic observations and simulations to advance the understanding of ET. At the observation side, the efforts include remote sensing signal retrieval (Mu et al., 2011; Zhang et al., 2010), flux tower network development, and data assimilation (Munier et al., 2015; Pan & Wood, 2006; Rodell et al., 2015). Meanwhile, the land surface modeling community has developed many numerical models that include ET simulation with different process representations, parameterizations, data requirements, and model structures, such as the Global Land-Atmosphere Coupling Experiment (Koster et al., 2004) and Land Data Assimilation System (Rodell et al., 2004). The observations and numerical simulations play complementary and interdependent roles in advancing our knowledge about the various hydrological processes and systems. Hydroclimatic observations provide inputs and validation references for numerical models; meanwhile, models generate data with continuous space and time dimensions, which are often used for interpolating point-scale observation (Jung et al., 2009), observation network design, and conceptual validation (Pan et al., 2011). Moreover, observations also serve as the source for hydrologic concept development and hypothesis testing, such as the Budyko hypothesis on long-term ET (Budyko, 1974), the complementary relationship between actual and potential ET (Brutsaert & Stricker, 1979), and the evaporative fraction between latent heat flux and available energy (Shuttleworth et al., 1989). In turn, improved theory guides new observation acquisition and advances model improvement (Gulden et al., 2007). This paper assesses hydrologic data, model, and theory congruence with a focus on ET in the CONUS.

Although advances have been made in monitoring and simulating ET over several decades, there is a pressing need to systematically evaluate observation and model consistency and enhance their complementary outputs for hydrologic knowledge discovery (Montanari et al., 2013; Shuttleworth, 2007; Sivapalan et al., 2011). Hydrologists nowadays often face a paradoxical situation: large amounts of data exist yet data uncertainty is inadequately assessed. Therefore, when the modeling community addresses the sensitivity of model performance to forcing data or parameters (Badgley et al., 2015; Montanari & Di Baldassarre, 2013; Xia, Peter-Lidard, et al., 2015), conclusions made on model evaluation (Cai et al., 2014; Swenson & Lawrence, 2015; Xia et al., 2016; Xia, Hobbins, et al., 2015) are essentially conditioned on the quality of reference observation data. Due to potential errors with a reference observation, a small discrepancy between model output and the reference may not necessarily mean that the model is acceptable; meanwhile, a poor fit to a set of noisy observation data does not provide a sufficient reason to reject a model. The efforts in reducing the discrepancy between model outputs and observations in model calibration exercises may fail to improve the model and result in a set of overconfident parameters if the reference observations involve systematic errors (Hejazi & Cai, 2009). Thus mistakes, such as accepting a wrong model or rejecting a good model, can be made due to unreliable reference data. Model evaluation can be further complicated when multiple inconsistent reference observations are available. For example, Cai et al. (2014) reported a reasonably good agreement in ET annual mean estimates between simulations from land surface models (LSMs) in Phase 2 of the North American Land Data Assimilation System (NLDAS-2) and two remote sensing ET products (Jung et al., 2009; Mu et al., 2011). However, Xia et al. (2016) found that the same LSMs failed in generating the ET seasonal cycle observed from gridded FLUXNET observations (Jung et al., 2009).

It has been argued that the model evaluation process should be *diagnostic*, that is, to obtain knowledge that can be used to either validate or reject the hypotheses underlying the model conceptualization and structure, which can eventually lead to improved models and advanced theories (Gupta et al., 2008). Hydrologic responses simulated by a model can rarely capture the full spectrum of hydrologic dynamics and/or hydrologic variability (Kumar, 2015) that, however, can be reflected by observations. Especially, current data acquisition has gone beyond what some existing LSMs can take as inputs. New variables, such as terrestrial water storage (TWS; Long et al., 2015), are now available at the global scale. Xia et al. (2017) evaluated the monthly TWS anomaly and the individual water storage components from three LSMs (i.e., Community Land Model version 4.0, NOAH-MP and Catchment Land Surface Model Fortuna 2.5, all including a groundwater component) against Gravity Recovery and Climate Experiment (GRACE). However, the change of TWS, which is widely caused by human interferences, remains as an outstanding issue with hydrologic modeling in general since few models have a reasonable depiction of the human dimension and its interactions with hydrologic processes (Vogel et al., 2015).

To deal with the situations described above, a hydrologic theory that represents falsifiable conceptualization of the real world is needed to diagnose the biases or errors involved in either observation or model, or both. Hydrologic theories can play a key role in bridging the gaps between models and observations, synthesizing our understanding of hydrologic phenomena and expanding hydrologic knowledge (Clark et al., 2016; Kirchner, 2006). For example, water balance is usually used as a closure constraint for multivariable observations (Gao, Tang, Ferguson, et al., 2010; Sheffield et al., 2009); the Budyko-type water-energy coupling relationship is applied to assessing the ET average and interannual variability in the International Satellite Land Surface Climatology Project Initiative (Koster et al., 2006). However, compared with the progresses in data and model development, theory development is limited in hydrology. Therefore, developing new theories and making better use of existing theories to underpin current models and data are urgently needed (Beven, 2012; Clark et al., 2016; Kirchner, 2006).

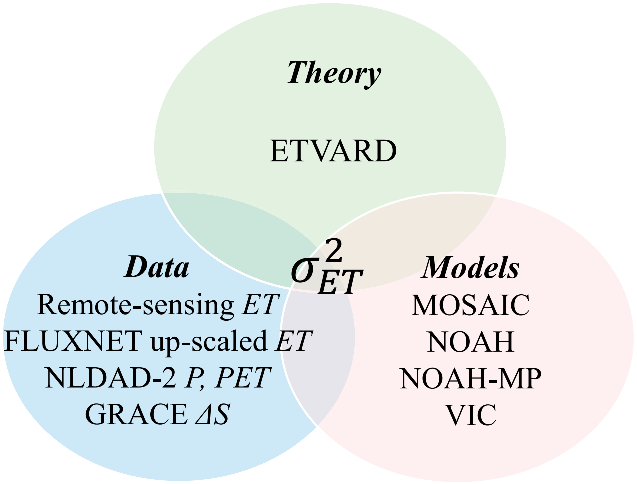

Figure 1. The congruence among hydrologic theories, multisource multivariable hydroclimatic observation data, and multiple numerical models represent our organized understanding of hydrologic processes. ETVARD = Evapotranspiration Temporal VARiance Decomposition; ET = evapotranspiration; NLDAS-2 = Phase 2 of the North American Land Data Assimilation System; PET = potential evaporation; GRACE = Gravity Recovery and Climate Experiment.

In this study, we treat an LSM as a set of hypotheses (Beven, 2012; Clark et al., 2011, 2015b, 2016) that are posed for the various hydrologic processes such as runoff generation, infiltration, and ET. For each of those processes, there can be several alternative hypothetical forms to describe the process; for example, infiltration can be described by Richard's equation or Green-Ampt method; potential ET can be calculated by Penman-Monteith equation or temperature-based methods. Since these processes are interconnected, the hypotheses should be tested in a systematic framework. For example, Koster and Suarez (1999) related ET variance to precipiation variance based on the Budyko theory and evaluated ET simulations from LSMs at the river basin scale. Recent research efforts have been made to develop frameworks for LSMs intercomparison, diagnosis, and benchmarking in land surface modeling communities and set model evaluation in a hypothesis test framework. Such efforts include the Framework for Understanding Structural Errors (Clark et al., 2008), the Joint U.K. Land Environment Simulator (Best et al., 2011), the Structure for Unifying Multiple Modeling Alternatives (Clark et al., 2015a, 2015b) and the PALS Land Surface Model Benchmarking Evaluation Project (Best et al., 2015). These comprehensive frameworks examine the simulation of relevant hydroclimatic and land surface processes (e.g., ET, infiltration, and streamflow) through intercomparison of the various model configurations and process representations (e.g., VIC calculates ET from soil evaporation, canopy evaporation, and vegetation transpiration (Gao, Tang, Shi, et al., 2010), and NOAH calculates ET from snow sublimation, bare soil evaporation, canopy water evaporation, and vegetation transpiration (Niu et al., 2011).

This paper presents a diagnostic framework based on an observation-model-theory triplet (Figure 1) to examine both the congruence and discrepancy of ET temporal variance from observations, models, and theories and provide guidelines for observation and model improvement. We adopt the Evapotranspiration Temporal VARiance Decomposition (ETVARD) framework provided by Zeng and Cai (2015) as a theoretical diagnostic tool. The original Budyko theory used in many previous studies as a constraining relationship in assessing hydrologic variable variability is usually suitable for long-term averages of the hydrologic variables, assuming that the long-term watershed system storage remains stable. This assumption is invalid for assessing variability at a relatively short time scale (annual or monthly) and for watersheds with systematic terrestrial storage change over a long-term period. ETVARD, based on an extension of the Budyko theory, takes into consideration of TWS change and quantitative attributes to the sources of ET variance to climatic and hydrologic components (Zeng & Cai, 2015). Taking ETVARD as a theory, five experiments are designed to assess the congruence among theory, data, and model, as well as the gaps between multiple hydrologic observations and LSMs on their estimate of ET temporal variability. The five experiments focus on ET monthly variance in the following aspects: (1) quantifying the climatic and hydrologic components of ET variance, such as precipitation (*P*), potential evaporation (*PET*), and TWS change (Δ*S*) to find the controlling factors on ET variance; (2) assessing the consistency of ET variance from four observation-based products and their bias/uncertainty compared to ETVARD framework and the compatiblity of a set of multivariable (i.e., *P*, PET, ET, and Δ*S*) observations under the theoretical ETVARD framework with differnet observation-based ET products; (3) cross evaluating of the four LSMs (MOSAIC, NOAH, and VIC from NLDAS-2 project (Mitchell et al., 2004; Xia, Mitchell, Ek, Cosgrove, et al., 2012; Xia, Hobbins, et al., 2015) and NOAH-MP (Cai et al., 2014)) subject to multisource reference observations to show how reference data sets affect the model evaluation; (4) benchmarking the four LSMs with ETVARD and diagnosing the possible deficits in each of the LSMs; and (5) evaluating the LSMs' simulation of the TWS to GRACE-estimated storage and their effects on the hydrologic system components of σ_{ET}. Based on the results of the five experiments, we will address the following questions: What hypotheses can be made from the diagnosis of the observation-model-theory triplet? What will be the major factor to ET variance (climatic variables versus TWS) in a particular region? What possible bias and uncertainty are involved in the various observation based ET products? Which aspects of LSMs and in which regions should be improved for more accurate simulation of intra-annual variance of ET?

2. Methodology, Data Sources, and Models

2.1. ET Temporal Variance Decomposition and Data Sources

Introducing TWS change as a variable in watershed water balance over a period, Zeng and Cai (2015) extended the Budyko relationship and, based on which, established an equation to decompose ET temporal variance into multiple contributing components as shown below:

$$\sigma_{\mathrm{ET}}^2 = w_{\mathrm{P}}\sigma_{\mathrm{P}}^2 + w_{\mathrm{PET}}\sigma_{\mathrm{PET}}^2 + w_{\Delta S}\sigma_{\Delta S}^2 + w_{\mathrm{P,PET}}\,\mathrm{cov}_{\mathrm{P,PET}} + w_{\mathrm{P},\Delta S}\,\mathrm{cov}_{\mathrm{P},\Delta S} + w_{\mathrm{PET},\Delta S}\,\mathrm{cov}_{\mathrm{PET},\Delta S} \quad (1)$$

where σ represents the standard deviation, cov represents the covariance, and w represents the weighting factors, which quantify the contribution from different variance/covariance sources to ET variance. The weighting factors, as shown in equation S1 in the supporting information, are calculated from the aridity index ($\overline{\phi} = \overline{\mathrm{PET}}/\overline{\mathrm{P}}$), Budyko equation $F\left(\overline{\phi}\right)$ (Yang et al., 2008), and its first-order derivative $F'\left(\overline{\phi}\right)$, which is also detailed in Zeng and Cai (2016). By equation (1), ET variance is determined by long-term climatic condition (through the weighting factors), climatic fluctuations (σ_{P}^2 and σ_{PET}^2) and their phasing ($\mathrm{cov}_{\mathrm{P,PET}}$), hydrologic storage variability ($\sigma_{\Delta S}^2$), and its response to climate ($\mathrm{cov}_{\mathrm{P},\Delta S}$ and $\mathrm{cov}_{\mathrm{PET},\Delta S}$). ETVARD provides an analytic way to decompose ET variance into climatic and hydrologic components and offers an independent estimate of ET variance based solely on hydroclimatic and catchment storage data (see Figure 2).

We can further aggregate the ET variance components based on their sources into two categories: One represents the contribution to ET variance from the variability of climatic forcing (σ_{ETF}^2) and the other from hydrologic storage (σ_{ETS}^2), that is,

$$\sigma_{\mathrm{ETF}}^2 = w_{\mathrm{P}}\sigma_{\mathrm{P}}^2 + w_{\mathrm{PET}}\sigma_{\mathrm{PET}}^2 + w_{\mathrm{P,PET}}\,\mathrm{cov}_{\mathrm{P,PET}} \quad (2)$$

$$\sigma_{\mathrm{ETS}}^2 = w_{\Delta S}\sigma_{\Delta S}^2 + w_{\mathrm{P},\Delta S}\,\mathrm{cov}_{\mathrm{P},\Delta S} + w_{\mathrm{PET},\Delta S}\,\mathrm{cov}_{\mathrm{PET},\Delta S} \quad (3)$$

Equation (1) can then be written as follows:

$$\sigma_{\mathrm{ET}}^2 = \sigma_{\mathrm{ETF}}^2 + \sigma_{\mathrm{ETS}}^2 \quad (4)$$

The time scale of ET variance depends on the time scale of the various variance/covariance terms. This study addresses ET variance at the monthly scale, while the analysis on ET variance at both annual and monthly scale can be found in Zeng and Cai (2016).

By assessing σ_{ET}^2, σ_{ETF}^2, and σ_{ETS}^2 by cells in the CONUS, the spatial patterns of the ET temporal variance are calculated. The assessments based on ETVARD will be used as a reference, and those from multiple ET products are compared to the reference, by which the possible bias and uncertainty involved in each of the ET products and their spatial patterns will be discussed.

This study uses monthly meteorological forcing data (P and PET) obtained from the NLDAS-2 (Mitchell et al., 2004; Xia, Mitchell, Ek, Sheffield, et al., 2012). P in NLDAS-2 is a product of gauge-only National Oceanic and Atmospheric Administration Climate Prediction Center, which conducted orographic adjustment of daily precipitation based on the Parameter-elevation Relationships on Independent Slopes Model (PRISM) climatology. The nonprecipitation land-surface forcing fields for NLDAS-2 are derived from the analysis fields of the National Centers for Environmental Prediction North American Regional Reanalysis and further vertically adjusted to account for the vertical difference between the North American Regional Reanalysis and NLDAS fields of terrain height. PET is calculated from modified Penman scheme (Mahrt & Ek, 1984) from the land-surface forcing fields for NLDAS-2. More details about the NLDAS-2 forcing data can be found at http://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php. The same forcing data fields are also used to drive the NLDAS-2 LSMs. P and PET from NLDAS-2 forcing data sets have a spatial resolution of 0.125° by 0.125° and cover the period from 1979 to 2015.

The terrestrial water storage (TWS) measured by the twin GRACE satellites is based on the distance change between the two satellites due to gravity field variation (Tapley et al., 2004). GRACE satellites primarily capture the mass change caused by TWS since other temporal changes of mass are negligible. GRACE-based TWS includes the sum of storage in various media such as aquifer, soil profile, snow/glacier, and surface reservoir/lake. The GRACE satellites provide a unique measurement of TWS with a large spatial coverage (Lettenmaier & Famiglietti, 2006) and have been widely applied for hydrologic studies such as groundwater depletion assessment (Famiglietti et al., 2011), water budget closure estimation (Pan et al., 2011), and LSM improvement (Gulden et al., 2007) and evaluation (Cai et al., 2014; Xia et al., 2017). The GRACE TWS data set used in this study, with a spatial resolution at 1° by 1°, provides a monthly time series from January 2003 to June 2013. The monthly terrestrial storage change (ΔS) is calculated as the difference from the monthly GRACE TWS time series (Landerer & Swenson, 2012) based on the Center for Space Research Release 5.0 at the University of Texas at Austin (http://www.csr.utexas.edu/).

2.2. Multisource Multivariable Hydroclimatic Observations and ET Products Based on Observations

Four ET products, three based on remote sensing observation and one based on FLUXNET, are assessed in this study. The remote sensing product by Tang et al. (2009) calculates ET as the combination of bare soil evaporation and vegetation transpiration based on the constant daily evaporative fraction assumption (Shuttleworth et al., 1989). Bare soil evaporation is estimated from surface radiation budget and soil temperature, and vegetation transpiration is calculated using the complementary relationship to bridge the actual ET with potential evaporation calculated by the Priestley-Taylor scheme. The data set covers the extent of the CONUS from 2001 to 2008 at a spatial resolution of 0.05° by 0.05° and is denoted as $ET_{RS\ -\ UW}$ in this study. The data and more details about the methodology can be found at Evaporation Estimation Using Remote Sensing at the University of Washington. This ET product has been used to assess watershed water budget (Gao, Tang, Ferguson, et al., 2010) and ET interannual variability (Cheng et al., 2011).

Another ET product by Mu et al. (2011) calculates ET from vegetation transpiration and soil evaporation based on the Penman-Monteith scheme presented in Mu et al. (2007). Vegetation evaporation is further separated into wet canopy surface evaporation and dry canopy vegetation transpiration, and the rates are regulated by aerodynamics resistance and surface resistance. Soil evaporation is divided into saturated soil potential evaporation and moist soil evaporation constrained by soil moisture stress. The monthly version global ET from 2000 to 2009 at 0.5° by 0.5° spatial resolution is obtained from MOD16 Global Terrestrial Evapotranspiration Data Set (http://www.ntsg.umt.edu/project/modis/mod16.php) and denoted as $ET_{RS\ -\ MOD16}$ in this study. The ET product has been applied for many studies such as drought assessment (Mu et al., 2013) and LSM evaluation (Cai et al., 2014).

The third ET product is from Global Land Evaporation Amsterdam Model (GLEAM) by Martens et al. (2017). The GLEAM uses the Priestley-Taylor equation to calculate the potential evaporation and builds a multilayer water balance model to determine the water stress. Semiempirical relationships between soil moisture and stress for evaporation and the depths of root zone depend on the land cover categories including bare soil, low vegetation, and tall vegetation. The daily actual ET from 1987 to 2017 at 0.25° by 0.25° spatial resolution is obtained from GLEAM v3.2a (https://www.gleam.eu/#downloads) and denoted as $\mathrm{ET}_{\mathrm{GLEAM}}$ in this study. The forcing data of GLEAM v3.2a include reanalysis net radiation and air temperature, a combination of gauge-based reanalysis and satellite-based precipitation, reanalysis and satellite-based soil moisture, and satellite-based vegetation optical depth (Martens et al., 2017).

The ET product developed by Jung et al. (2009) is different from the remote sensing products in terms of both data sources and retrieval algorithms. It is essentially the spatial upscaling of point measurements from eddy covariance flux tower. This approach uses model tree ensemble, a machine learning technique, to upscale current global network of eddy covariance towers (FLUXNET) and evaluates results from the *virtual reality* produced by Lund-Potsdam-Jena managed Land biosphere model simulation. This ET estimate has been applied for ET trend analysis (Jung et al., 2010) and LSM improvement (Bonan et al., 2011) and evaluation (Cai et al., 2014). The global monthly ET estimate from 1982 to 2008 at 0.5° by 0.5° spatial resolution is denoted as $ET_{FLUXNET}$ in this study.

Since NLDAS-2 meteorological forcing, GRACE TWS, and the ET products have different spatial resolutions and temporal coverage, these data sets (as summarized Table 1) are processed to calculate ET variance using the following procedures: First, the time series from all data sets are spatially aggregated and matched at the 1° by 1° GRACE cells to for the CONUS domain of latitude between 25°N and 53°N and longitude between 67°W and 125°W. Cells with missing or incomplete data from the observation products are excluded when aggregated to the 1° by 1° grid. At the 1° by 1° spatial resolution (the coarsest spatial resolution among

Table 1
Multisource Hydroclimatic Observations

Variables	Source	Spatial resolution	Temporal coverage
P	NLDAS-2	0.125° by 0.125°	1979–2015
PET	NLDAS-2	0.125° by 0.125°	1979–2015
ET	Tang et al. (2009)	0.05° by 0.05°	2001–2008
	Mu et al. (2011)	0.5° by 0.5°	2000–2009
	Jung et al. (2009)	0.5° by 0.5°	1982–2008
	Martens et al. (2017)	0.25° by 0.25°	1980–2017
Δ*S*	GRACE	1° by 1°	January 2003 to June 2013

Note. NLDAS-2 = Phase 2 of the North American Land Data Assimilation System; PET = potential evaporation; ET = evapotranspiration; GRACE = Gravity Recovery and Climate Experiment.

these data sets), we assume that the TWS change caused by lateral flow is negligible. Note that for the cases of long-distance water diversions, for example, California diverts water from the Colorado River using an aqueduct that spans over 300 km; the storage changes at some locations (cells) can be caused by those occurring at other locations. Second, the weighting factors in equation (1) are calculated from long-term average climate condition based on NLDAS-2 P and PET from 1979 to 2015. Third, climatic variabilities (i.e., σ_{P}, σ_{PET}, and $\text{cov}_{\text{P,PET}}$ in equation (2)) are calculated from monthly time series during 1979–2015 and the variabilities associated with storage change (i.e., $\sigma_{\Delta S}$, $\text{cov}_{\text{P},\Delta S}$,and $\text{cov}_{\text{PET},\Delta S}$ in equation (3)) are calculated from monthly time series during the period of January 2003 to June 2013 during which GRACE is available. ET variances from direct observations are calculated from their temporal coverages. Note that it is ideal to compare ET variance calculated from different data sets with the same temporal coverage. However, the period overlapped by all data sets in this study is only 6-year long (January 2003 to December 2008), which is too short to get an accurate and stable estimate of ET monthly variance. In order to use the longest records from each data set and make ET variance calculated from each data set comparable, we assume that the variance and covariance terms are statistically stationary during the whole period (January 1975 to January 2015). Based on our analysis, the average differences of ET variance calculated between complete and overlapping periods are less than 5 mm^2.

2.3. Multiple Operational LSMs

Four operational LSMs (MOSAIC, NOAH, and VIC NOAH-MP) use the same climate forcings and vegetation cover parameters at the same temporal and spatial scales, allowing the intercomparison to focus on model structure. This study uses monthly scale model inputs (i.e., P and PET) and outputs (i.e., ET and Δ*S*), which are available at the National Oceanic and Atmospheric Administration/ National Centers for Environmental Prediction/Environmental Modeling Center NLDAS ftp servers (http://www.emc.ncep.noaa.gov/mmb/nldas/). A LSM calculates terrestrial ET from soil, canopy, snow, and vegetation, depending on the processes formulated in the LSMs. TWS change (Δ*S*) includes the changes of soil moisture, snow, and aquifer storage. To compare a LSM to ETVARD, the LSM results obtained at the resolution of 0.125° by 0.125° (the common scale used by all the LSMs) must be aggregated to 1° by 1°, as the resolution of GRACE data. Each of the LSMs simulates PET with different methods but using the same meteorological forcings, while ETVARD uses the PET that is also calculated using NLDAS-2 forcing data (Mahrt & Ek, 1984).

2.4. Experiment Design

1. Experiment 1: We set $\sigma^2_{\text{ETVARD}-\text{GRACE}}=\sigma^2_{\text{ETF}}+\sigma^2_{\text{ETS}-\text{GRACE}}$ using P and PET observations from NLDAS-2 and GRACE-based Δ*S* (Figure 2). Through this experiment, we will investigate the climatic and hydrologic contributions in σ^2_{ET} and their spatial patterns for the CONUS. Since GRACE satellites capture TWS change from soil, snow, groundwater, river channels, and lakes, and they respond to climate fluctuation differently, this experiment can identify the dominant control on ET variance at different hydroclimatic settings. We further interpret the σ^2_{ET} components by ETVARD and assess the relative role of climatic variables and hydrologic system variables (including human interferences).
2. Experiment 2: We use $\sigma^2_{\text{ETVARD}-\text{GRACE}}$ from Experiment 1 as a reference to compare $\sigma^2_{\text{ET}-\text{Obs}}$ from the four observation-based ET products ($\sigma^2_{\text{RSUW,}}$ $\sigma^2_{\text{RSMOD16}}$, $\sigma^2_{\text{FLUXNET}}$, and σ^2_{GLEAM}). We compare the four estimates to $\sigma^2_{\text{ETVARD}-\text{GRACE}}$ at particular locations of interest and their spatial distribution across the CONUS.
3. Experiment 3: We compare the ET variance from the four LSMs (denoted as $\sigma^2_{\text{ET}-\text{LSM}}$) to that from the four observation-based ET products $\sigma^2_{\text{ET}-\text{Obs}}$ in Experiment 2. By calculating the difference of σ^2_{ET} between each LSM simulation and each observation-based product, we will obtain a matrix showing the comparisons of the LSMs and the observations. This experiment is designed to show how the conclusion of model evaluation varies with the references.
4. Experiment 4: We calculate $\sigma^2_{\text{ETVARD}-\text{LSM}}$ from ETVARD using the same climate forcings (i.e., P and PET) as those used in the LSMs and the terrestrial storage Δ*S* simulated by the four LSMs. Depending on the LSMs configuration and model structure, Δ*S* is summed up from soil profile, snow water equivalent, and

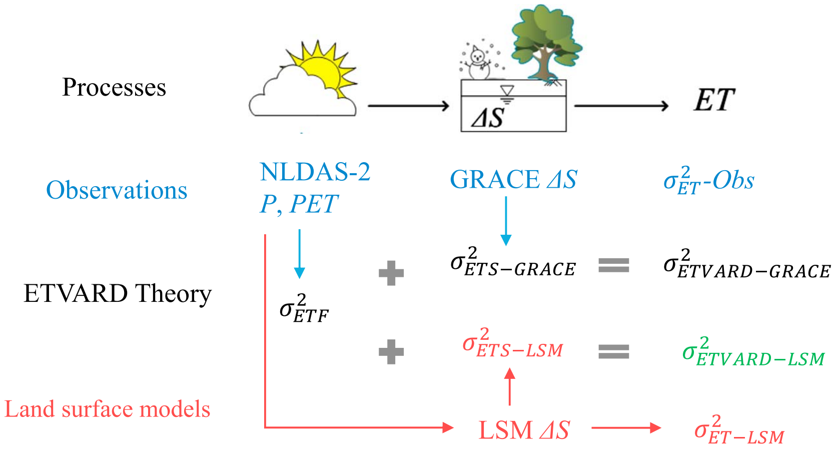

Figure 2. Schematics of hydrologic processes along with various ET variance estimates and its components from observation, simulation and ETVARD approaches. The σ^2_{ETF} is the climatic components of ET variance; σ^2_{ETS} is the storage components of ET variance, which can be calculated from GRACE-observed or LSM-simulated ΔS; $\sigma^2_{\mathrm{ET-Obs}}$ includes four observation based products; $\sigma^2_{\mathrm{ET-LSM}}$ is calculated from four LSM simulation time series. Since the climatic components (driving force) in $\sigma^2_{\mathrm{ETVARD-GRACE}}$ and $\sigma^2_{\mathrm{ETVARD-LSM}}$ are the same, the differences between them are due to model differences in storage change. ETVARD = Evapotranspiration Temporal VARiance Decomposition; ET = evapotranspiration; NLDAS-2 = Phase 2 of the North American Land Data Assimilation System; PET = potential evaporation; GRACE = Gravity Recovery and Climate Experiment; ETS = evapotranspiration from hydrologic storage; ETF = evapotranspiration of climatic forcing; LSM = land surface model.

groundwater storage (only available in NOAP-MP model). Note that different LSMs have different soil layers and depths each soil layer, and we sum all soil layers in each model. Thus, in this experiment, $\sigma^2_{\mathrm{ET-LSM}}$ and $\sigma^2_{\mathrm{ETVARD-LSM}}$ are based on the *same* climatic and storage change inputs for each LSM model. Thus, this experiment isolates the effects on σ_{ET} estimates associated with processes represented by LSMs from that associated with the input data. Therefore, the comparison will focus on the difference caused by the model structure (i.e., physical process representation) of an LSM and the analytical form of ETVARD.

5. Experiment 5: We assess the impact on σ^2_{ETS} from the TWS change (ΔS) estimates based on two sources: GRACE-based observation ($\sigma^2_{\mathrm{ETS-GRACE}}$) and LSM-simulation ($\sigma^2_{\mathrm{ETS-LSM}}$), that is, the only variable of interest in Experiment 5 is ΔS. With the same climatic forcings (i.e., P and PET) for ETVARD and LSMs, we focus on the comparison of the hydrologic component σ^2_{ETS} (equation (3)). σ^2_{ETS} includes water storage change variability, the correlation between P and ΔS (e.g., soil moisture replenishment, aquifer recharge due to rainfall excess, and pumping and water withdrawal in dry days) and the correlation between PET and ΔS (e.g., snow melting and thaw). Note that σ^2_{ETS} represents the water storage related components in σ^2_{ET} and therefore can be negative or positive. The TWS in GRACE observation includes groundwater and surface water storage (e.g., lakes, reservoirs and river channel storage), which are generally not simulated by operational LSMs yet (Xia et al., 2017). Through Experiment 5, we expect to identify locations for LSM improvement regarding the interaction between land surface processes and groundwater, by natural processes (e.g., groundwater recharge/discharge and snow dynamics), human activities (e.g., groundwater pumping), or both.

3. Results

3.1. Experiment 1: σ^2_{ET} Components in the CONUS

Figures 3a–3c display the magnitudes and spatial distribution of σ^2_{ET} components from climatic variables P, PET, and their phase, respectively. As can be seen from Figure 3a, the contribution from P ($w_{\mathrm{P}}\sigma^2_{\mathrm{P}}$) is more than 2,000 mm^2 in California and southern Florida. In the High Plains, $w_{\mathrm{P}}\sigma^2_{\mathrm{P}}$ is also notable (around 1,500 mm^2) and decreases gradually from south to north. $w_{\mathrm{P}}\sigma^2_{\mathrm{P}}$ is small (less than 500 mm^2) in the Mountain States and negligible above the Great Lakes (since ET from the water surface is not constrained by fluctuation in P). The contribution from PET variability ($w_{\mathrm{PET}}\sigma^2_{\mathrm{PET}}$) is relatively small compared to $w_{\mathrm{P}}\sigma^2_{\mathrm{P}}$ and exhibits a sharp contrast

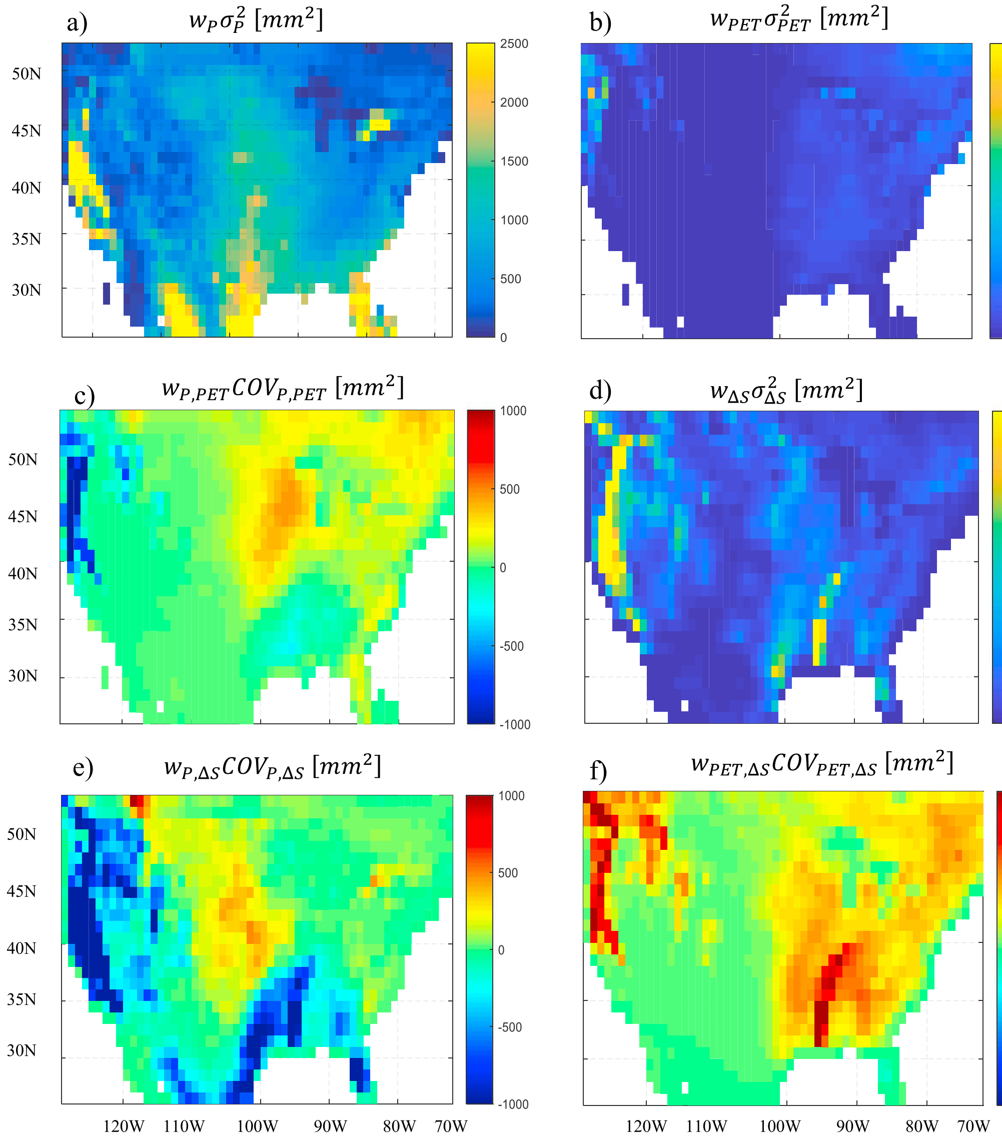
a) $w_P\sigma_P^2\ [mm^2]$
b) $w_{PET}\sigma_{PET}^2\ [mm^2]$
c) $w_{P,PET}COV_{P,PET}\ [mm^2]$
d) $w_{\Delta S}\sigma_{\Delta S}^2\ [mm^2]$
e) $w_{P,\Delta S}COV_{P,\Delta S}\ [mm^2]$
f) $w_{PET,\Delta S}COV_{PET,\Delta S}\ [mm^2]$
50N
45N
40N
35N
30N
120W
110W
100W
90W
80W
70W
2500
2000
1500
1000
500
0
1000
500
0
-500
-1000

Figure 3. (a–f) Climate and storage components of σ_{ET}^2 derived from ETVARD in equation (1) based on P and PET from NLDAS-2 forcing and ΔS from GRACE. ETVARD = Evapotranspiration Temporal VARiance Decomposition; PET = potential evaporation; NLDAS-2 = Phase 2 of the North American Land Data Assimilation System; GRACE = Gravity Recovery and Climate Experiment.

along the east-west direction, as shown in Figure 3b. The Mountain States (west of 100th Meridian West, commonly considered as arid and semiarid region) have negligible $w_{\mathrm{PET}}\sigma_{\mathrm{PET}}^2$. The coastal regions of Washington and Oregon states, Northern California, and the Great Lakes have significant amount of contribution from PET variability (more than 500 mm^2). In these regions, ET is limited by energy supply, and fluctuation in PET propagates to ET variance. In addition, the northeastern region has a noticeable $w_{\mathrm{PET}}\sigma_{\mathrm{PET}}^2$ component (between 200 and 400 mm^2). Figure 3c shows that the in-phase of P~PET enhances ET variance in the Corn Belt, while the out-of-phase P~PET reduces ET variance in the coastal regions of Washington, Oregon, and California Central Valley due to the Mediterranean climate in those regions (as shown in Figure S1). As ET is constrained by water availability during the dry season and by energy availability during the wet season, ET intraannual fluctuation is dampened by the climate pattern.

Figures 3d–3f display the magnitudes and spatial distribution of the contribution to σ_{ET}^2 from the variance of ΔS, the covariance of P and ΔS, and the covariance of PET and ΔS, respectively. The $w_{\Delta s}\sigma_{\Delta S}^2$ is more than

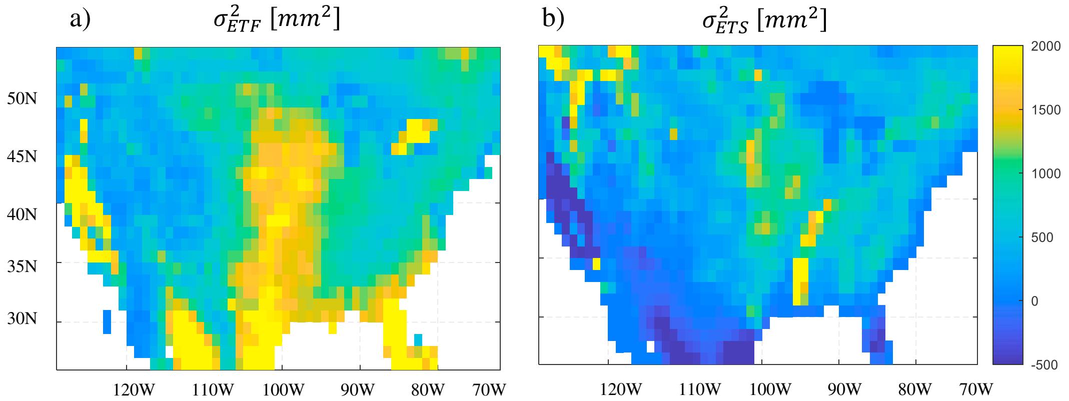

Figure 4. (a) Climatic (σ^2_{ETF} in equation (2)) and (b) storage (σ^2_{ETS} in equation (3)) components of σ^2_{ET}.

1,000 mm^2 in the Pacific Northwest and California and more than 500 mm^2 in the south part of the High Plains and Mississippi Embayment region. As can be seen from Figure 3e, the interaction between P and ΔS ($w_{\mathrm{P},\Delta s}\mathrm{cov}_{\mathrm{P},\Delta S}$) significantly reduces σ^2_{ET} in the western coast especially in California (2,000 mm^2) but slightly enhances σ^2_{ET} (around 500 mm^2) in the North High Plains and part of the East. Although $\mathrm{cov}_{\mathrm{PET},\Delta S}$ is significant in the West (see Figure S1f in the supporting information), its contribution to ET variance concentrates to a limited region in California due to a low weighting factor shown in Figure S2f. The South and the Appalachian Mountains also have a fairly significant $w_{\mathrm{PET},\Delta S}\mathrm{cov}_{\mathrm{PET},\Delta S}$ component (more than 300 mm^2).

Adding the climatic components together by equation (2), σ^2_{ETF}, the overall ET variance from the climate variables is shown in Figure 4a. Generally, the distribution of σ^2_{ETF} follows that of precipitation in most places. In the Corn Belt, the in-phase of P~PET provides a favorable condition for crop water consumption, yielding a relatively large σ^2_{ETF} (more than 1,500 mm^2). The coastal regions in Washington and Oregon have relatively mild σ^2_{ETF}, since the out-of-phase P~PET in those regions dampens ET variance. The Appalachian Mountains have low σ^2_{ETF} (less than 1,000 mm^2); the Mountain States have the lowest σ^2_{ETF} (less than 500 mm^2) in magnitude.

The aggregated hydrologic system components of ET variance σ^2_{ETS} by equation (3) is shown in Figure 4b. Note that σ^2_{ETS}, denoting the responses of TWS change to climate and human interferences on σ_{ET}, can be negative (i.e., a dampening effect) or positive (i.e., an enhancing effect). In general, the magnitudes of σ^2_{ETS} are smaller than those of σ^2_{ETF}, indicating the major impact of climatic variance in general. However, σ^2_{ETS} enhances ET variance (more than 1,000 mm^2) over the High Plains and Mississippi downstream and reduces ET variance (more than 500 mm^2) in California Central Valley. These regions with strong σ^2_{ETS} components overlap with major aquifers that have been depleted for irrigation (Konikow, 2015). Anthropogenically induced storage change either enhances (in the High Plains) or dampens (in California) ET variance through the covariance between catchment water storage and precipitation seasonality as shown in Figure 3f. The Cascade Range and northern part of the Rocky Mountains also have positive σ^2_{ETS}, mainly because the snow accumulating and melting processes provide a temporal redistribution of water from cold to warm seasons.

3.2. Experiment 2: Multisource σ^2_{ET} Comparison

Since the total ET variance is all positive, the following analysis on ET variance from multiple observations is assessed in terms of ET standard deviation. In Figure 5a, σ_{ETVARD} ranges between 0 and 60 mm; the maximum σ_{ETVARD} occurs across the High Plains and decreases toward the west, with the minimum located along the east of Sierra Nevada Mountains. Florida also has noticeable σ_{ETVARD} (above 40 mm); the western coastal region and the Appalachian-Northeast line also have moderate σ_{ETVARD} (around 30 mm). As shown in Figure 5b, the remote sensing σ_{RSUW} exhibits similar spatial zonation to σ_{ETVARD}, with the peak value in the Midwest and the coastal region of the North Pacific. The σ_{RSUW} is generally larger than 40 mm on other

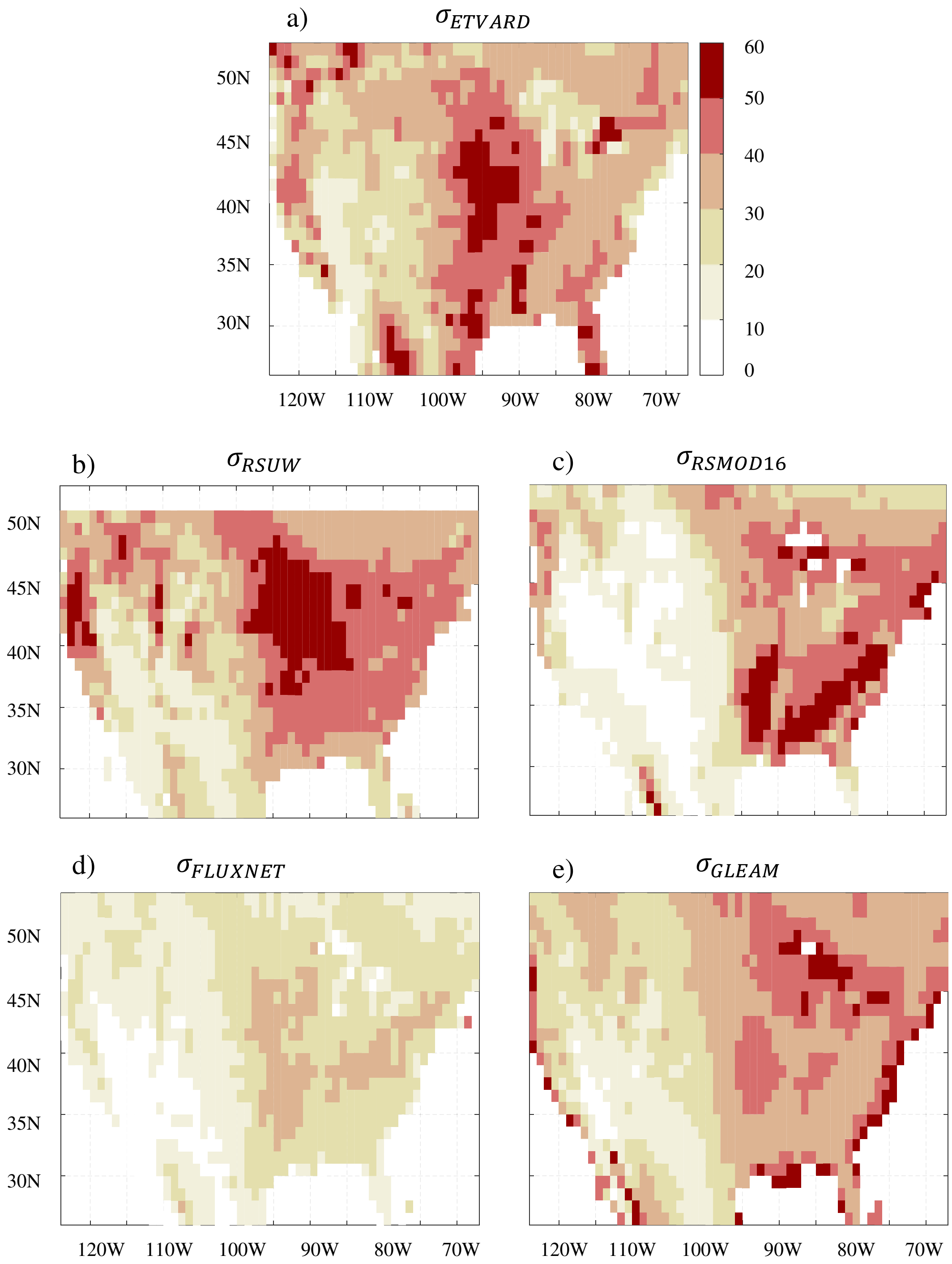

Figure 5. Spatial pattern of ET standard deviation (mm) estimated by four observation-based products: (a) ETVARD, (b) ET_{RSUW}, (c) $ET_{RSMOD16}$, (d) $ET_{FLUXNET}$, and (e) ET_{GLEAM}. ET = evapotranspiration; ETVARD = Evapotranspiration Temporal VARiance Decomposition; RSUW = ET from the University of Washington; RSMOD16 = from MOD16 Global Terrestrial ET Data Set; GLEAM = from Global Land Evaporation Amsterdam Model.

parts of the East and less than 30 mm in the Mountain States, with the minimum located along the east of Sierra Nevada Mountains. Remote sensing $\sigma_{RSMOD16}$ shows a contrasting west-east spatial pattern (Figure 5c). The $\sigma_{RSMOD16}$ in western CONUS is mostly below 20 mm and is smaller than that from σ_{ETVARD} or σ_{RSUW} (above 20 mm in the region). However, the northern Pacific Coast is exceptional with $\sigma_{RSMOD16}$

around 30 mm. The $\sigma_{\mathrm{RSMOD16}}$ is about 10 mm smaller than σ_{ETVARD}, σ_{RSUW}, and σ_{GLEAM} in the western CONUS. The peak values (larger than 50 mm) of $\sigma_{\mathrm{RSMOD16}}$ are located along the downstream of the Mississippi River and the Southeast. The Midwest and Northeast has moderate $\sigma_{\mathrm{RSMOD16}}$ between 30 and 50 mm. The FLUXNET upscaling estimate $\sigma_{\mathrm{FLUXNET}}$ is shown in Figure 5d. The $\sigma_{\mathrm{FLUXNET}}$, ranging between 0 and 40 mm, is smaller than the other four estimations. The spatial distribution of maximun $\sigma_{\mathrm{FLUXNET}}$ is similar to that of σ_{ETVARD}, extending from the Midwest to the south part of the High Plains. The Appalachian-Northeast line also has substantial $\sigma_{\mathrm{FLUXNET}}$; the western CONUS has $\sigma_{\mathrm{FLUXNET}}$ generally below 20 mm, which shows a similar range to that of $\sigma_{\mathrm{RSMOD16}}$. The σ_{GLEAM} in Figure 5e ranges between 10 and 50 mm and exhibits an east-west gradient. The σ_{GLEAM} is significant in the Midwest. In addition, GLEAM produces highest σ_{ET} over the Great Lakes and less than 50 mm in other regions. Being opposite to $\sigma_{\mathrm{RSMOD16}}$, σ_{GLEAM} shows lower values (below 40 mm) along the Appalachian Mountain and Lower Mississippi River. Similar to the spatial pattern in the western CONUS by $\sigma_{\mathrm{RSMOD16}}$ and $\sigma_{\mathrm{FLUXNET}}$, GLEAM produces slightly higher σ_{ET} between 10 and 30 mm. It is noted that GLEAM produces erroneously high σ_{ET} along the coastline, probably due to a mixed signal from the sea and land.

It is not surprising to see the discrepency of spatial patterns of ET variance from these five estimates, but it is difficult to draw the conclusion on which product is more reliable than others, since the true value is not known. In general, σ_{ETVARD} and σ_{RSUW}, the two independent estimates yield similar spatial patterns and magnitudes. The $\sigma_{\mathrm{FLUXNET}}$ seems to be underestimated, compared to other four products. Probably, the flux tower sites are too sparse to capture the heterogeneity of ET for a large region. Errors in σ_{ETVARD} may exist at coastal cells where GRACE-estimated ΔS contains signals of sea water.

The frequency histograms of the residuals between σ_{ETVARD} and σ_{RSUW}, $\sigma_{\mathrm{RSMOD16}}$, $\sigma_{\mathrm{FLUXNET}}$, or σ_{GLEAM} are plotted in Figure 6. As shown in Figure 6a, the residual between σ_{RSUW} and σ_{ETVARD} fits a Gaussian distribution with a mean of 0.52 mm and standard deviation of 11.09 mm. The small residual (i.e., σ_{RSUW} - σ_{ETVARD}) indicates that this set of multivariable hydroclimatic observations (i.e., NLDAS-2 P and PET and GRACE-estimated ΔS) are statistically unbiased relative to σ_{RSUW} under the general laws embedded in ETVARD.

The residual between $\sigma_{\mathrm{RSMOD16}}$ and σ_{ETVARD} as plotted in Figure 6b yields a slightly bimodal distribution, and a Gaussian fit results in a mean of -7.80 mm and standard deviation of 14.59 mm, the largest uncertainty among the four observation-based ET products. The residual between $\sigma_{\mathrm{FLUXNET}}$ and σ_{ETVARD} in Figure 6c yields a Gaussian distribution with mean of -15.31 mm and standard deviation of 8.15 mm. The relatively small residual standard deviation indicates ET_{FLUXNET} may have relatively small uncertainty than the other three ET products, while the large residual mean indicates that $\sigma_{\mathrm{FLUXNET}}$ is probably underestimated when using ETVARD as a benchmark. The histogram of residual between σ_{GLEAM} and σ_{ETVARD} in Figure 6d fits a Gaussian distribution with a mean of -2.27 mm and standard deviation of 12.82 mm, similar to that by $\mathrm{ET}_{\mathrm{RSUW}}$. The cells along the coastline contribute to residuals larger than 20 mm.

3.3. Intercomparison of σ_{ET} Among Multiple Reference Observations, ETVARD, and Multiple LSMs (Experiment 3)

The monthly σ_{ET} from the four LSMs ranges from 0 to 60 mm as shown in Figure 7. The four LSMs commonly produce high σ_{ET} (above 40 mm) in Midwest and low σ_{ET} (below 20 mm) in the region west of meridian 100°W. Meanwhile, the four LSMs produce different levels of σ_{ET} in the northeastern region of CONUS, where σ_{ET} is above 30 mm for MOSAIC and NOAH-MP and around 20 mm for NOAH and VIC. The LSM-simulated σ_{ET} values show significant differences along the West Coast compared to the four observation-based estimates in Figure 5. Compared to four observation-based estimates, which all yield noticeable σ_{ET} (larger than 30 mm) along the West Coast, though varying in magnitude, the four LSMs result in low σ_{ET} (20 mm) along the West Coast. Compared to the result of ETVARD, the four LSMs consistently generate low σ_{ET} in the West Coast as well. A unique contributor to σ_{ET} along the West Coast is the Mediterranean climate. By ETVARD, P and PET are out-of-phase between the rainfall season and the warm season, resulting in a negative climatic component (i.e., $w_{\mathrm{P,PET}}\mathrm{cov}_{\mathrm{P,PET}}$) in σ_{ET} in this region, as shown in Experiment 1 and Figure 3c. In addition, the contributions from TWS change in this region are also notable. In California, the TWS release during the dry season leads to a significant reduction in σ_{ET} via a negative $w_{\mathrm{P},\Delta S}\mathrm{cov}_{\mathrm{P},\Delta S}$ component, while snow melting during the warm season enhances σ_{ET} with a positive $w_{\mathrm{PET},\Delta S}\mathrm{cov}_{\mathrm{PET},\Delta S}$ component in the coastal region of Oregon and Washington. Thus, the relatively low σ_{ET} from the four LSMs in the West Coast is probably due to the

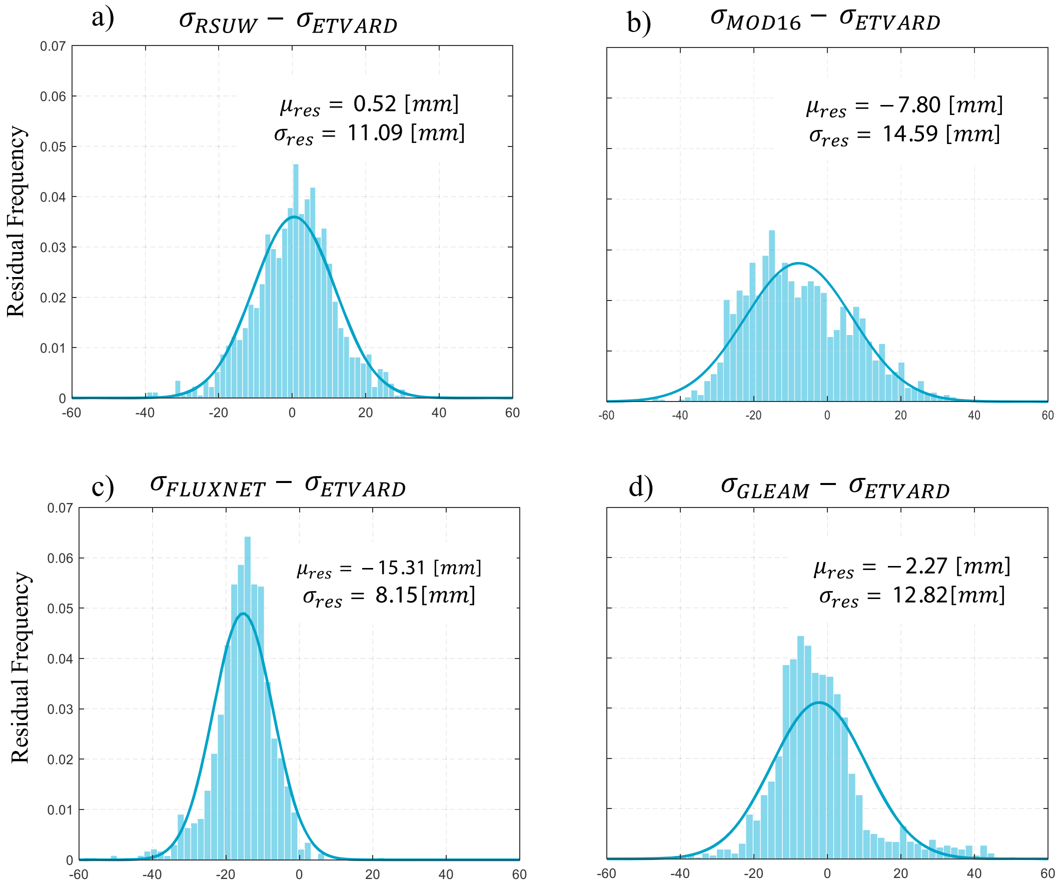

Figure 6. Residual histograms of total ET variance f between ETVARD and (a) ET_{RSUW}, (b) $ET_{RSMOD16}$, (c) $ET_{FLUXNET}$, and (d) ET_{GLEAM}. Residuals are fitted into normal distribution. ET = evapotranspiration; ETVARD = Evapotranspiration Temporal VARiance Decomposition; RSUW = ET from the University of Washington; RSMOD16 = from MOD16 Global Terrestrial ET Data Set; GLEAM = from Global Land Evaporation Amsterdam Model.

Mediterranean climate or the limited water storage representation in the models. More detailed results on terrestrial storage change effects should be referred to Experiment 5, which includes the assessment of GRACE-estimated terrestrial storage change on ET variance.

The average residual (σ_{ETres}) between σ_{ET} from a LSM and that from an observation-based product is calculated as the mean absolute difference between the four observation products over all cells in the CONUS, that is, $\sigma_{\text{ETres}} = \frac{1}{n}\sum_{i=1}^{n}|\sigma_{\text{ETLSM}} - \sigma_{\text{ETObs}}|$. The pairwise intercomparisons are shown in Table 2. The σ_{ET} calculated by ETVARD is also used as a reference together with the observations. By each column of Table 2, one observation is used as the reference, and the model with the smallest residual is picked as the *best model*. For example, when σ_{ET} from ETVARD is treated as reference, MOSAIC model has the smallest residual (i.e., 8.41 mm) among the four models and is therefore chosen as the best model. It is surprising to find that each of the LSMs is identified once as the best model with the various references. This illustrates that intercomparison of the multiple LSMs is observation dependent. Recognizing the possible limitations of using any single model for problem solution, ensemble-based approaches have been widely used to handle model uncertainties, in which the results from the various models are combined with a certain given set of priorities (often subjective) on the models.

3.4. Model Processes Representation Assessment Using ETVARD as a Benchmark for LSMs (Experiment 4)

The σ_{ET} by ETVARD with ΔS inputs from each LSM is shown in Figure 8. The four σ_{ET} estimates exhibit a clear contrast along the east-west direction near the meridian 100°W line. For all cases, σ_{ET} from ETVARD using ΔS

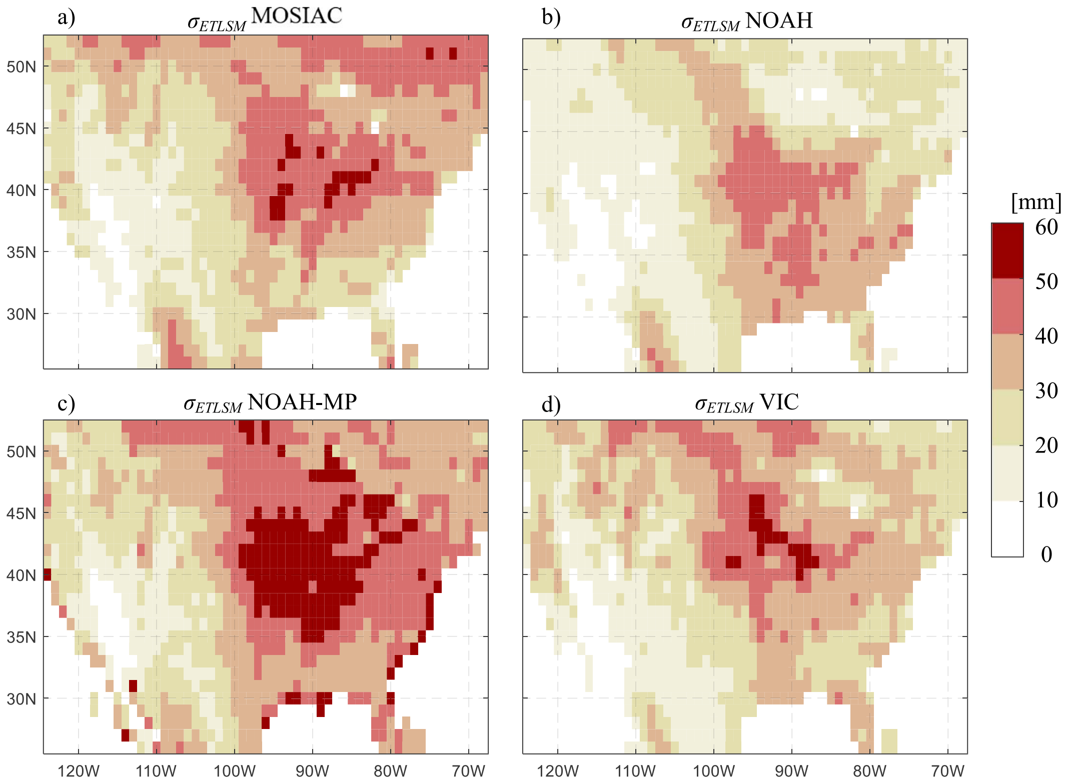

Figure 7. The σ_{ET} (i.e., the red σ_{ETLSM} in Figure 2) simulated by the four LSMs: (a) MOSAIC, (b) NOAH, (c) NOAH-MP, and (d) VIC, driven by the same forcing data sets and calculated at the same temporal and spatial resolution. ET = evapotranspiration; LSM = land surface model.

from all the LSMs is less than 20 mm in the west mountains and larger than 40 mm in the West Coast of California. The maximum σ_{ET} (about 50 mm) is generally located near the Midwest, while σ_{ET} with ΔS from NOAH-MP generates high σ_{ET} in the whole eastern United States except for areas along the Appalachian Mountains.

Figure 9 displays the differences in σ_{ET} from each of the four LSM results (i.e., $\sigma_{\mathrm{ETLSM}} = f_{\mathrm{LSM}}(\mathrm{P}, \mathrm{PET}, \Delta S_{\mathrm{LSM}})$ in Figure 2), where f_{LSM} represents a LSM model function) and ETVARD with ΔS simulation from each of the four LSMs as input (i.e., $\sigma_{\mathrm{ETVARD}} = f_{\mathrm{ETVARD}}(\mathrm{P}, \mathrm{PET}, \Delta S_{\mathrm{LSM}})$ in Figure 2), where f_{ETVARD} represents equation (1). Note that the inputs (P, PET, ΔS_{LSM}) to ETVARD and the LSMs are the same. Thus, Figure 9 isolates out the impact on σ_{ET} from the input data and explicitly shows the difference between an LSM and ETVARD caused by the physical process representation of σ_{ET} in LSM (i.e., f_{LSM}) and the analytical ETVARD (i.e., f_{ETVARD}). A common spatial pattern shared by the four LSMs is that σ_{ET} along the West Coast is significantly smaller (about 20 mm) than that from ETVARD. As discussed in Experiment 3, the most apparent σ_{ET} difference between LSM results and the observation-based estimates in Experiment 2 is also located along the West Coast. We have suggested that the difference may be caused by inaccurate simulation of TWS or by inadequate process representation under the Mediterranean climate. Xia et al. (2016) pointed out that some LSMs failed to acceptably simulate the annual cycle of the monthly mean ET in the Mediterranean climate. Some LSMs predict peak ET in spring, while others predict peak ET in summer. Since Mediterranean climate is water limited or energy limited during different seasons, the vegetation response to climate in these LSMs may not be well represented. For example, Cai et al. (2014)

Table 2
Pairwise σ_{ET} Differences Between Land Surface Models and Observation Products

Land surface models (mm)	ETVARD	RS-UW	MOD16	FLUXNET	GLEAM
MOSAIC	*8.41*	7.72	10.96	12.26	7.74
NOAH	12.03	11.96	10.52	*6.21*	8.29
NOAH-MP	8.95	*6.58*	11.16	15.49	*7.40*
VIC	9.61	8.52	*10.26*	9.12	8.65
Best model	MOSAIC	NOAH-MP	VIC	NOAH	NOAH-MP

Note. Columnwise comparison represents the average σ_{ET} residual when an observation-based σ_{ET} is used as reference, so the smallest absolute value in the column (in italic) indicates the best model. ETVARD = Evapotranspiration Temporal VARiance Decomposition; GLEAM = Gravity Recovery and Climate Experiment.

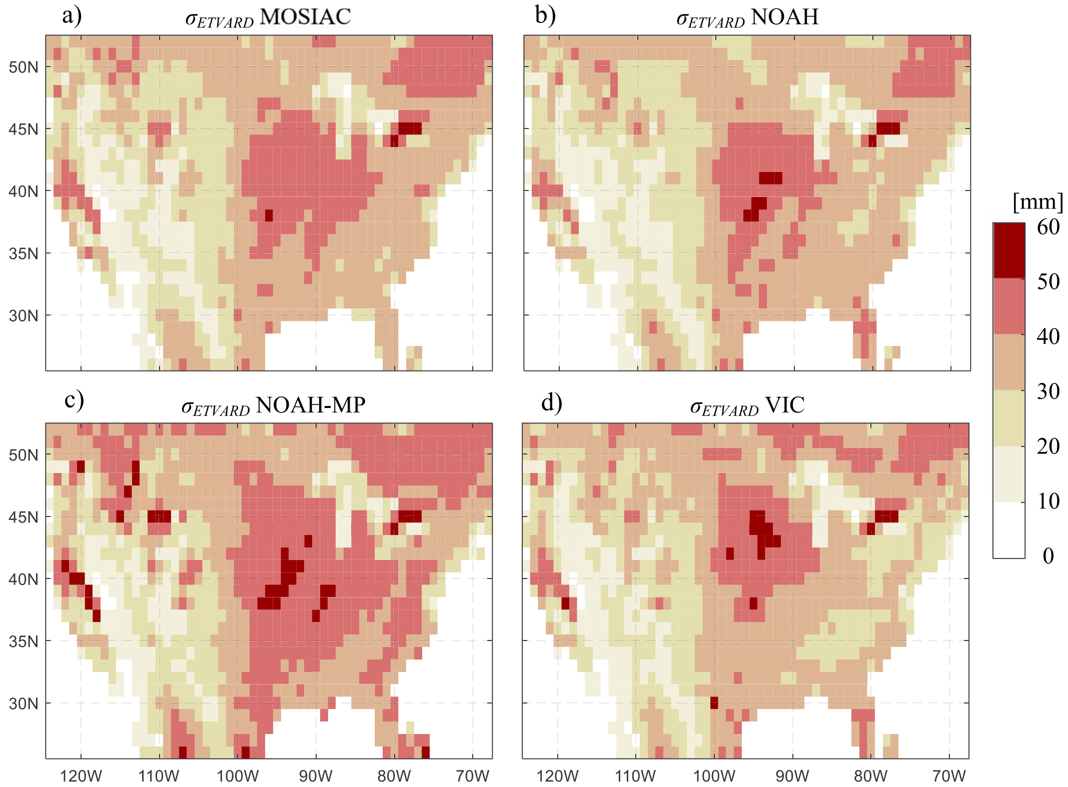

Figure 8. The σ_{ET} calculated by ETVARD (i.e., the green σ_{ETLSM} in Figure 2) with terrestrial water storage change (ΔS, including soil moisture, snow and/or groundwater) simulated by four LSMs: (a) MOSAIC, (b) NOAH, (c) NOAH-MP, and (d) VIC. ETVARD = Evapotranspiration Temporal VARiance Decomposition.

suggested that using a dynamic leaf phenology model in LSM (predicting LAI as a function of light, temperature, and soil moisture) improves the ET simulation using the prescribed monthly LAI. In a study that also discusses the discrepancies between LSMs and remote sensing observations of ET, Castle et al. (2016) showed that ET was underestimated by some LSMs during peak irrigation times. In this experiment that compares the differences in model processes representation of ET variance, we may further claim that the differences are mainly attributed to the representation of vegetation responses to water and energy stresses in dry and wet seasons, respectively.

Although σ_{ET} differences between LSMs and ETVARD are found in other regions, they are not consistently shared by the four LSMs. For instance, MOSAIC, NOAH-MP, and VIC generally yield slightly higher σ_{ET} (less than 5 mm) than that from ETVARD in the Midwest and Northeast, while NOAH exhibits higher σ_{ET} mainly in the Southeast. NOAH and NOAH-MP predict significant lower σ_{ET} (more than 20 mm) than ETVARD near Idaho, where the covariance between ΔS and PET contributes considerably to σ_{ET} in Figure 3f. This implies that the differences might be mainly associated with snow processes or vegetation's responses to solar radiation in NOAH and NOAH-MP. In addition, NOAH and VIC show significantly lower σ_{ET} than ETVARD and observation-based σ_{ET} around the southern region along meridian 100°W, where P is the dominant component in σ_{ET} as shown in Figure 3a.

Although we do not claim that any of the estimates by LSMs, ETVARD, or observation-based estimates are accurate, this experiment shows that in most of the regions in the CONUS the estimates from ETVARD and observations are more similar to one another in the West Coast and upper plains, as shown by the differences between Figures 5, 7, and 8. Following the analysis of the contribution sources of σ_{ET} in Experiment 1, we can target particular processes contributing to the disagreements for further studies. Moreover, taking the ETVARD as a benchmark, Experiments 1 and 4 can be used for identifying the processes controlling σ_{ET} and their spatial locations in the four LSMs. For example, Experiment 1 shows that

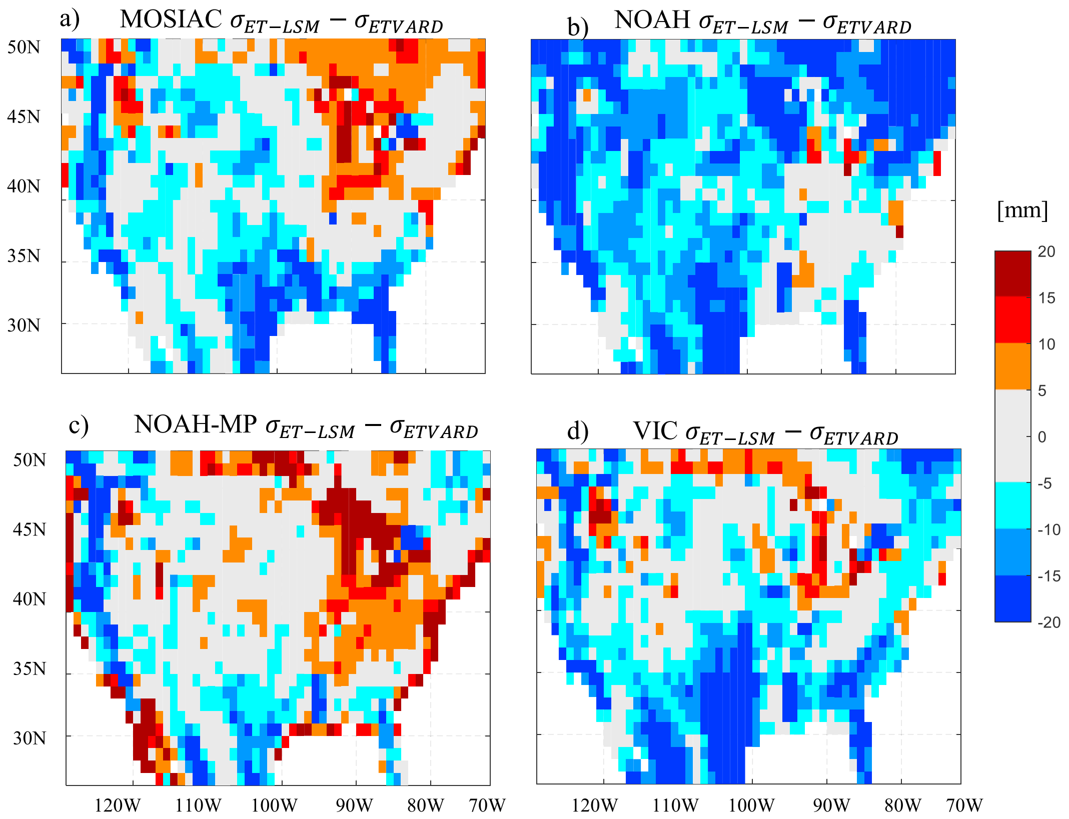
a) MOSIAC $\sigma_{ET-LSM} - \sigma_{ETVARD}$
b) NOAH $\sigma_{ET-LSM} - \sigma_{ETVARD}$
c) NOAH-MP $\sigma_{ET-LSM} - \sigma_{ETVARD}$
d) VIC $\sigma_{ET-LSM} - \sigma_{ETVARD}$
50N
45N
40N
35N
30N
120W
110W
100W
90W
80W
70W
[mm]
20
15
10
5
0
-5
-10
-15
-20

Figure 9. The σ_{ET} residual between ETVARD ($\sigma_{\mathrm{ET}} = f_{\mathrm{ETVARD}}(\mathrm{P}, \mathrm{PET}, \Delta S_{\mathrm{LSM}})$) and LSM ($\sigma_{\mathrm{ET}} = f_{\mathrm{LSM}}(\mathrm{P}, \mathrm{PET}, \Delta S_{\mathrm{LSM}})$), that is, f_{ETVARD} - f_{LSM}, with the same input data; this residual shows the pairwise discrepancy between the benchmarking ETVARD and aggregated processes in each of the four LSMs: (a) MOSAIC, (b) NOAH, (c) NOAH-MP, and (d) VIC.

the energy budget dominates σ_{ET} on the coast of Washington and Oregon. Therefore, energy-related processes such as snow dynamics or vegetation water demand should be examined in the models for these regions.

3.5. LSM Diagnosis Using TWS Observations (Experiment 5)

The terrestrial storage component of ET variance, σ^2_{ETS}, calculated from $\Delta S_{\mathrm{GRACE}}$ and four LSMs simulated ΔS_{LSM}, respectively, ranges from -800 to 1200 mm^2, as shown in Figure 10. The estimates from the four LSMs and GRACE are quite consistent in the South and the West, where ΔS buffers the ET fluctuation. In Idaho, all five σ^2_{ETS} estimates consistently indicate that ΔS enhances σ_{ET}, mainly due to the snow storage. The NOAH-MP exhibits high σ^2_{ETS} (larger than 1,000 mm^2) in a slightly larger area than other models. Experiment 4 shows that the snow processes (variance of the storage) or vegetation's response to solar radiation (via the covariance between ΔS and PET) in NOAH and NOAH-MP may be responsible for the difference between LSMs and ETVARD. Experiment 5 further finds that the vegetation's response to solar radiation (the covariance item, $w_{\mathrm{PET},\Delta S}\mathrm{cov}_{\mathrm{PET},\Delta S}$) can be the primary reason for the difference.

The most apparent σ^2_{ETS} difference between GRACE observation and LSM simulation appears in the Midwest and the High Plains. The $\sigma^2_{\mathrm{ETS-GRACE}}$ shows that ΔS substantially enhances σ_{ET} in the Midwest and the High Plains, while the four LSMs generate large σ^2_{ETS} generally to the east of meridian 90°W. The impacts on ET from agricultural practices and groundwater-based irrigation in these regions have been well recognized by both remote sensing estimates (Mutiibwa & Irmak, 2013; Strassberg et al., 2009) and groundwater well measurements (Haacker et al., 2015; McGuire, 2012). However, an accurate representation of heavily managed agricultural land use still remains a challenge in LSM formulation. LSMs generally have a relatively shallow soil profile (e.g., 2 m in VIC; Liang et al., 1994) which can be sufficient to characterize natural vegetation root water uptaking but cannot catch the effect of groundwater pumping, which decreases water storage deep in the aquifer,

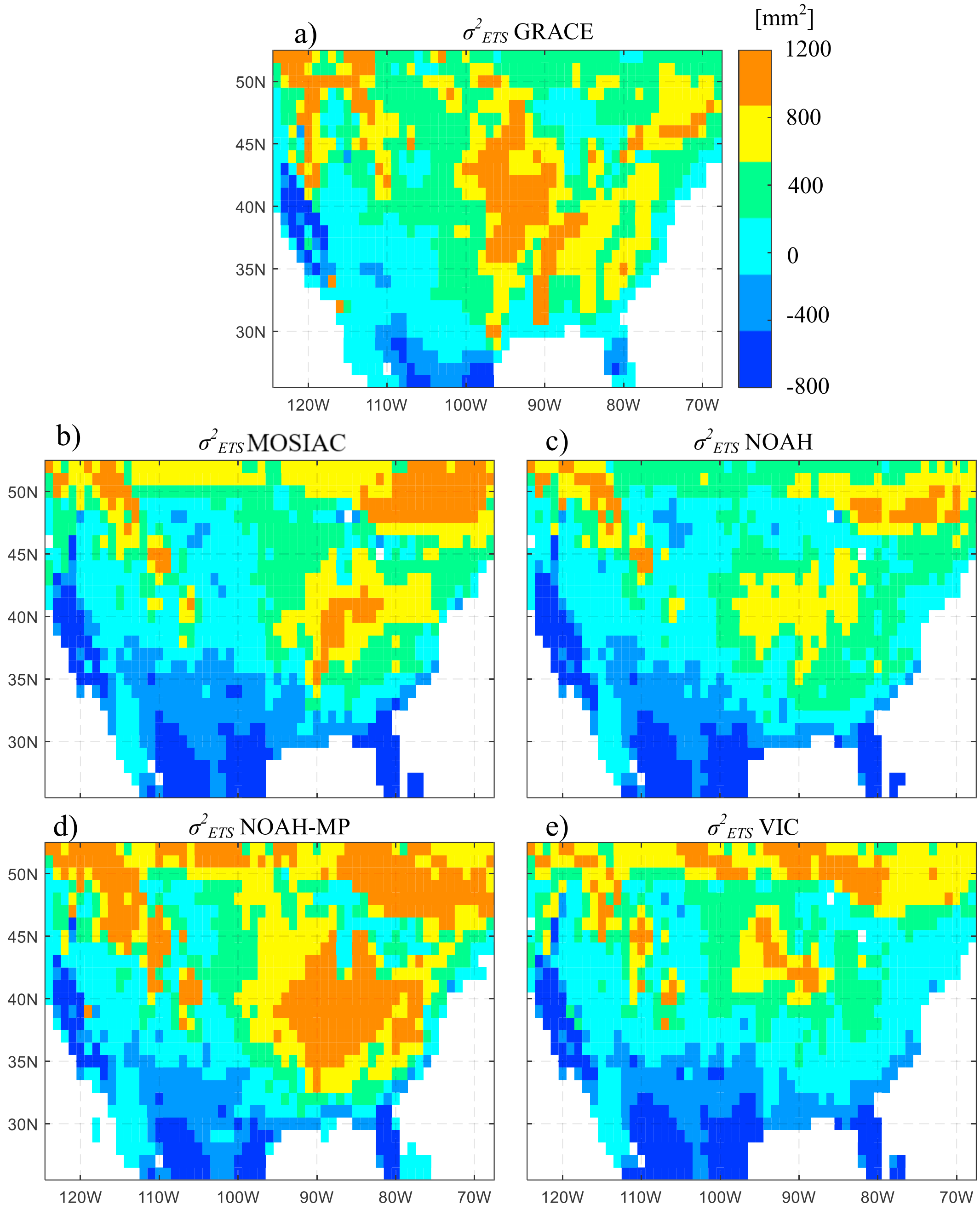

Figure 10. The σ^2_{ETS}, the terrestrial water storage change components in σ^2_{ET}, with Δ*S* from (a) GRACE observation and the four LSM simulations: (b) MOSAIC, (c) NOAH, (d) NOAH-MP, and (e) VIC. GRACE = Gravity Recovery and Climate Experiment; LSM = land surface model.

and in turn cannot reflect the effect of accumulated depletion of aquifer storage (Zeng & Cai, 2014). Although NOAH-MP has a simple aquifer representation, the transient decline in groundwater level results in a large amount of storage change which is beyond the storage capacity specified in LSMs. Thus, Experiment 5 unveils how a better simulation of Δ*S*, especially in intensively managed agricultural land, would improve the simulation of ET and ET variance in LSMs. It is noted that some cells along the Mississippi River in Figure 10a have high σ^2_{ETS}. This may not correctly reflect the storage components in ET variance, since the GRACE data show very high TWS variation amplitude due to the water fluctuation in the river channel (Cai et al., 2014).

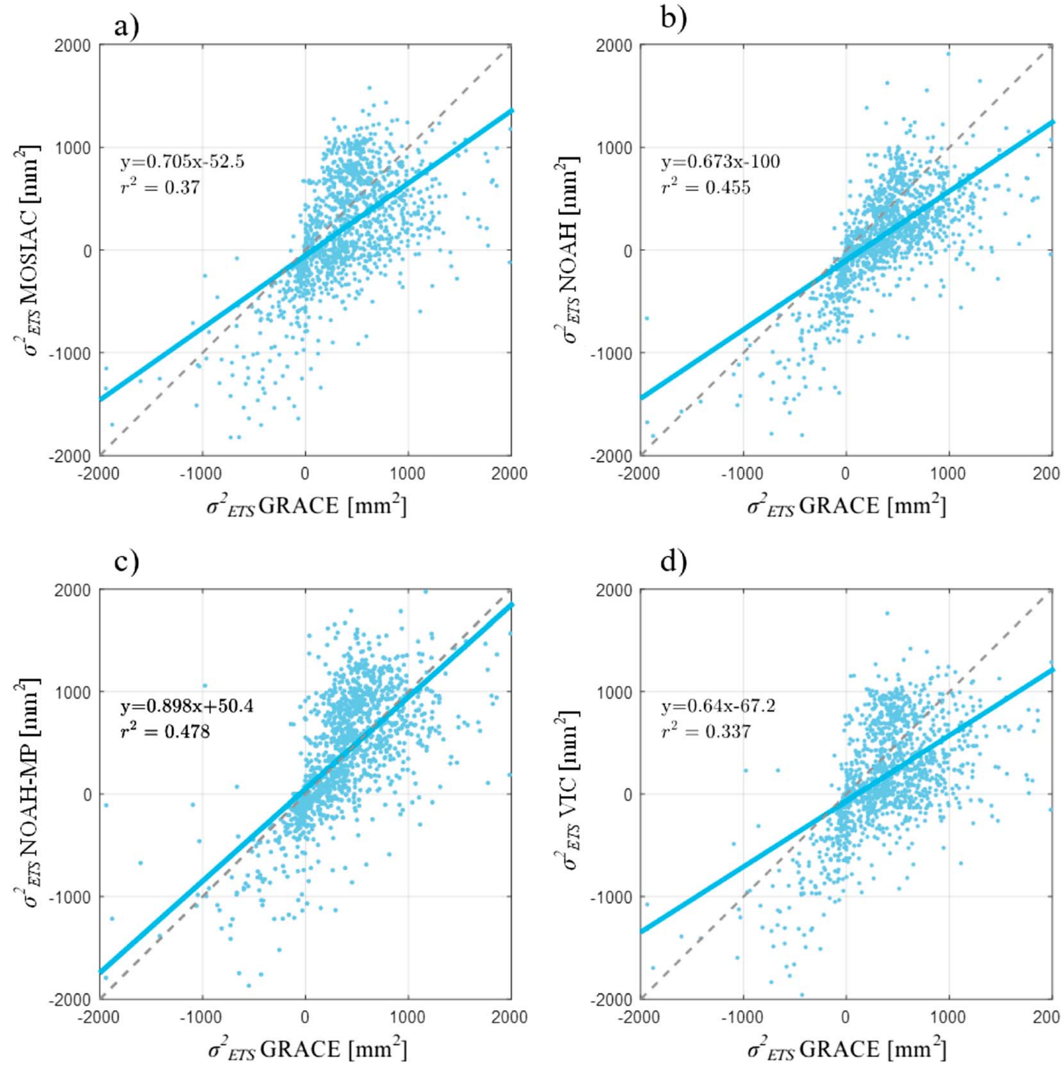

Figure 11. The scatter plot of σ^2_{ETS} from GRACE observation and the four LSM simulations: (a) MOSAIC, (b) NOAH, (c) NOAH-MP, and (d) VIC. The positive σ^2_{ETS} indicates cells where σ_{ET} is enhanced by terrestrial water storage change, and negative σ^2_{ETS} indicates cells where σ_{ET} is dampened by terrestrial water storage change.

The scatter plot of $\sigma^2_{\mathrm{ETS-GRACE}}$ and $\sigma^2_{\mathrm{ETS-LSM}}$ of the four LSMs in the CONUS is shown Figure 11. Overall, all LSMs yield smaller σ^2_{ETS} components than the GRACE-based σ^2_{ETS} (the regression slopes are less than 1). Among all the LSMs, NOAH-MP produces the closest σ^2_{ETS} to the GRACE-based estimate, which is probably due to the aquifer module (though a simple one) in the NOAH-MP. It is noted that in regions where ΔS buffers σ_{ET} (i.e., the $\sigma^2_{\mathrm{ETS}}<0$), the buffering effect by LSMs is consistently less than that reflected by GRACE observation. These regions are mainly located in the western mountainous regions where TWS plays a more dominant role in σ^2_{ETS} than other regions. Further study is needed to assess not only the accuracy of ΔS from GRACE but also the uncertainties involved in the LSMs, especially in the process representations associated with ΔS simulation.

4. Discussion

4.1. The Clustering of Most Important Components of ET Variance in Climatic Plane

Zeng and Cai (2015) qualitatively divided the $\left(\overline{\mathrm{P}}, \overline{\mathrm{PET}}\right)$ plane into several zones with various controlling factors on σ^2_{ET} based on the weighting factors. Here we take the largest absolute value of the six components in each

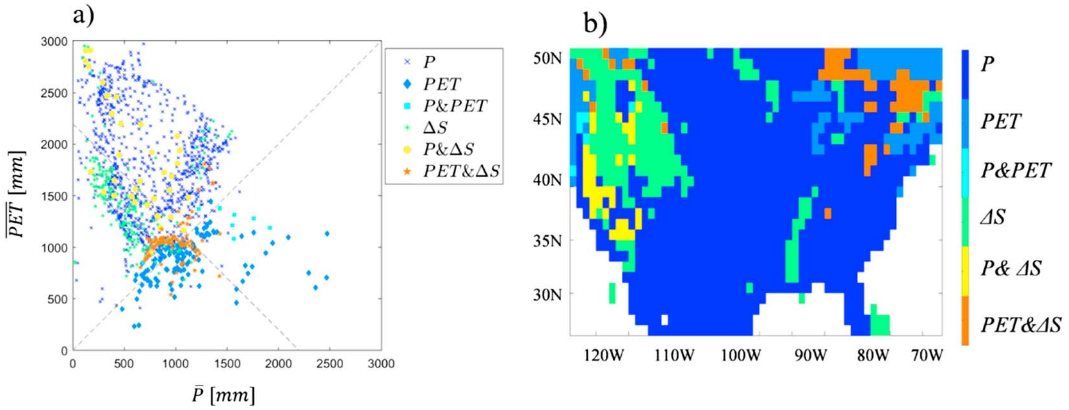

Figure 12. (a) Dominant σ^2_{ET} component in each grid in the CONUS in the $(\overline{\mathrm{P}}, \overline{\mathrm{PET}})$ plane. The 45° dashline ($\overline{\phi} = 1$) distinguish humid and arid climates. (b) Spatial distribution of dominant components of σ^2_{ET}. PET = potential evaporation.

grid (Figure 3) and identify that as the most important controlling component of ET variance. Those identifications are plotted in the $(\overline{\mathrm{P}}, \overline{\mathrm{PET}})$ plane as shown in Figure 12, which confirms that in the CONUS P and PET are the major controls of σ^2_{ET} in arid ($\phi > 1$) and humid regions ($\phi < 1$), respectively. The major components associated with ΔS are in the lower left region, where the water and energy fluxes have relatively small values (approximately, $\overline{\mathrm{P}} < 1{,}000$ mm and $\overline{\mathrm{P}} + \overline{\mathrm{PET}} < 2{,}200$ mm). Beyond these empirical thresholds, climate factors are the major components in σ^2_{ET} since the catchment storage has relatively limited capacity to buffer the water and energy fluctuations.

Exceptionally, several major components associated with storage (ΔS and $P\&\Delta S$) are far beyond a thresholds ($\overline{\mathrm{P}} + \overline{\mathrm{PET}} < 2,200$ mm) shown in Figure 12. These points represent the major components of California or the areas along the lower reaches of the Mississippi River as shown in Figure 12. The deviation of the dominant components in these regions may be due to the water storage change by agricultural water uses, which have significantly larger capability to use storage (e.g., pumping groundwater or surface water storage) than natural vegetation. This shows that human water use significantly affects ET, causing the dominant components of σ^2_{ET} to deviate from those in natural catchments. Another possible reason that storage change dominates the Mississippi River would be the large fluctuation in river channel storage change due to lateral flow, which is not considered in ETVARD framework. Thus, the $(\overline{\mathrm{P}}, \overline{\mathrm{PET}})$ plane provides a visual diagnostic tool to detect human interferences or possible errors on σ^2_{ET}.

The spatial pattern of σ^2_{ET} dominant components shows that P dominates σ^2_{ET} in most parts of the East and Central CONUS. PET dominates σ^2_{ET} in the Great Lakes and New England, since this region is strictly energy limited. The Mountain West shows much diversity on the dominant components. ΔS in Montana, Idaho, Northern Utah, and Northern Nevada dominates σ^2_{ET}, showing the important role of snow storage in local hydrologic dynamics. The cells in California show $P\&\Delta S$ as dominant factors, due to the coincidence between dry season and snow melting and groundwater use.

4.2. Implications of σ^2_{ET} Components for Model Development

The climatic components in equation (2) and storage components in equation (3) of σ^2_{ET} provide valuable information for hydrologic model development in terms of increasing the accuracy of model inputs and the improvement of model structures. For regions where σ^2_{ET} climatic components are significant (as shown in Figures 3a–3c and 4a), more reliable model input fluxes (i.e., P and PET) would improve the model performance. On the other hand, for example, σ^2_{ET} in the western CONUS is not significantly affected by PET (Figure 3b). Therefore, the hydroclimatic processes and models in this region may not need to be sensitive to the fluctuations in PET. Improving the model structure to better capture how the hydrologic state variable, S (e.g., snow, soil moisture, and groundwater), responds to climate is important in regions where σ^2_{ET} storage components are significant (Figures 3d–3f and 4b). For example, $w_{\mathrm{P},\Delta S}\mathrm{cov}_{\mathrm{P},\Delta S}$ represents catchments' response (both natural and anthropogenic) to P, such as groundwater recharge and pumping. Agricultural irrigation enhances the σ^2_{ET} in the High Plains and dampens the σ^2_{ET} in California (Figure 3e). LSMs do not capture these processes mainly due to two reasons. First, most LSMs do not have an aquifer storage component, and the buffering effect of soil profile is limited. Second, LSMs lack a good representation of farmers' water use behavior, such as irrigation. In some regions the groundwater table may be quite deep and not naturally coupled with land surface processes. However, farmers' pumping well can access the groundwater that cannot be utilized by natural vegetation, connecting the groundwater dynamics to the crop water consumption and climatic fluctuations. Therefore, farmers' irrigation behavior should be reasonably represented in the models developed for these regions. The $w_{\mathrm{PET},\Delta S}\mathrm{cov}_{\mathrm{PET},\Delta S}$ represents catchments' response to PET, such as snow melting and vegetation water demand. Figure 3f indicates that the snow dynamics in north Pacific Coast and vegetation dynamics in Eastern CONUS are important processes controlling the σ^2_{ET} in these regions, respectively.

4.3. Limitations and Future Perspectives

The purpose of this study is to utilize the theoretic ETVARD framework for the reconciliation between LSMs and observations focusing on ET temporal variance at the month scale. This study does not aim at providing a comprehensive framework for LSM diagnosis, as done by other efforts (Best et al., 2015; Clark et al., 2015a). However, we illustrate a meaningful framework in which ETVARD is used to disaggregate and diagnose σ_{ET} in LSMs while systematically adopting hydrologic observations that reflect some dynamics that may not be well captured by LSMs. We do not explicitly assess the impact of climatic forcings on σ_{ET}, given that the four LSMs underlying the three experiments use the same set of forcings (from NLDAS-2 project). If another set of climatic forcings (P and PET) are available, this study can be extended to account how different climatic forcings impact σ_{ET}.

With a growing amount of hydroclimatic observation data, LSMs are being improved. However, new theories and hypotheses are still needed to synthesize hydrologic knowledge through the observation-model-theory triplet. Researchers have recognized that the existing and even growing gap between models and theories is impeding the progress of hydrologic science (Clark et al., 2016). The ETVARD framework is our first attempt toward congruence among the observation-model-theory triplet.

Another issue is that existing hydrologic relationships are generally obtained in natural watersheds with minimal human interferences. Existing LSMs essentially simulate the virgin hydrologic cycling without fully considering anthropogenic impacts. As human activities play an increasing role in transforming hydrologic processes, such as irrigation and baseflow (Wang & Cai, 2009), hydrologic models should be developed or improved to better capture the anthropogenic components at multiple temporal and spatial scales (Vogel et al., 2015).

5. Conclusions

We illustrate how multisource, multivariable hydroclimatic observations, multiple LSMs, and ETVARD (a theoretical ET variance assessment framework) can serve complementarily by cross diagnosing each other, through five systematically designed experiments. We particularly show the role of ETVARD as an independent diagnosis tool in the observation-model-theory triplet. Based on σ^2_{ET} derived from ETVARD (Experiment 1), we characterize the spatial distribution of σ^2_{ET} and its climatic and hydrologic components across the CONUS. Although the contribution to σ^2_{ET} from climatic variables is larger than that from the hydrologic system variables in most of the regions of the CONUS, we identify some regions such as California and the lower reach of Mississippi River, where TWS-related components have significantly changed the σ^2_{ET}. In those regions, groundwater pumping for irrigation (e.g., in California) and water withdrawal from surface water (e.g., lower reach of Mississippi River) have led to systematic change of the terrestrial storage.

Based on the comparison of four observation-based ET products using ETVARD as a reference, we propose some diagnostic hypotheses regarding possible bias and uncertainty involved in the various ET products: $\mathrm{ET}_{\mathrm{RS-UW}}$ captures the high σ^2_{ET} signals in the Midwest, with negligible *bias* and moderate uncertainty over the CONUS; $\mathrm{ET}_{\mathrm{FLUX-MTE}}$ systematically underestimates σ^2_{ET} over the CONUS but with the lowest level of uncertainty; $\mathrm{ET}_{\mathrm{RS-MOD16}}$ has medium bias with the highest level of uncertainty, and the spatial distribution of high σ^2_{ET} signal from $\mathrm{ET}_{\mathrm{RS-MOD16}}$ is different from other estimates. Note that the reference value derived from ETVARD itself depends on the quality of the multiple data sources that are used to estimate the climatic and hydrologic variables involved in ETVARD (P and PET from NLDAS-2 and ΔS from GRACE), including errors that can be caused by the aggregation processes of the data sources with different spatial and temporal resolutions. This complexity encourages the use of a diagnostic framework as illustrated in this paper, which assumes uncertainties from estimates from a theoretical framework (ETVARD), observations, and models and attempt to identify congruence among the three.

Further, it is found that any of the four models compared can be the best one for a certain set of reference observations, which confirms our argument that intercomparison of multimodels depends on the reference observation. Therefore, simply minimizing the residual between model and observation may result in rejecting a good model with unreliable observation (the so-called Type I error) or accepting a wrong model with unreliable observation (Type II error). It is also found that σ_{ET} derived from ETVARD is consistently closer to observation-based estimates than the LSM simulations, especially in regions along the West Coast, Midwest, and High Plains. All four LSMs might underestimate σ_{ET} along the West Coast due to the Mediterranean climate and human water use; these models might also underestimate the terrestrial storage contribution to ET variance in the High Plains compared to the ETVARD estimate and GRACE observation. This is probably due to the inappropriate representation of groundwater pumping and its impact on ET and other hydrologic processes in those LSMs. Furthermore, compared to GRACE-based estimates, the four LSMs do not capture the high σ^2_{ETS} signal in the Midwest and High Plains. This is likely due to the limited representation of the hydrologic processes in the LSMs that control the terrestrial storage changes such as groundwater balance in aquifers and vegetation dynamics.

Overall, via five systematically designed experiments, we diagnose the congruence in σ_{ET} among multisource and multivariable hydrologic observations, multiple LSMs, and ETVARD. Each experiment independently and complementarily provides information for the various assessments. Given possible errors and uncertainties in multiple models and multiple observations, the observation-model-theory triplet with a theoretical diagnostic tool is useful for cross validating hydrologic theories, observations, and models. A match between observation and simulation does not necessarily capture the reality unless it is consistent with a confirmed, generic theory, that is, achieving congruence among observation, model, and theory. In particular, in this era with increasing multisource and multivariable hydrologic observations and improvement in various hydrologic models, we demonstrate the role of generic hydrologic theories (e.g., ETVARD in this study) as a bridge between models and observations and encourage further efforts along the line for the hydrologic community.

Acknowledgments
The authors appreciate the constructive comments by the Editor and three reviewers. All data used in this study are publicly available from the following sites: http://www.emc.ncep.noaa.gov/mmb/nldas/ (NLDAS-2 forcing and LSM simulations), https://www.gleam.eu/#downloads (GLEAM ET product), https://www.bgc-jena.mpg.de/bgi/index.php/Services/Overview (FLUXNET up-scaled ET product), http://www.ntsg.umt.edu/project/modis/mod16.php (MODIS 16 ET product), and http://www.csr.utexas.edu/ (GRACE terrestrial storage).

References

Badgley, G., Fisher, J. B., Jiménez, C., Tu, K. P., & Vinukollu, R. (2015). On uncertainty in global terrestrial evapotranspiration estimates from choice of input forcing datasets. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM-D-14-0040.1

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. . L. H., Ménard, C. B., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677–699. https://doi.org/10.5194/gmd-4-677-2011

Beven, K. (2012). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience*, *344*(2), 77–88. https://doi.org/10.1016/j.crte.2012.01.005

Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., et al. (2011). Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data. *Journal of Geophysical Research*, *116*, G02014. https://doi.org/10.1029/2010JG001593

Brutsaert, W., & Stricker, H. (1979). An advection-aridity approach to estimate actual regional evapotranspiration. *Water Resources Research*, *15*(2), 443–450. https://doi.org/10.1029/WR015i002p00443

Budyko, M. I. (1974). *Climate and Life*. San Diego, CA: Academic Press.

Cai, X., Yang, Z.-L., Xia, Y., Huang, M., Wei, H., Leung, L. R., & Ek, M. B. (2014). Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS test bed. *Journal of Geophysical Research: Atmospheres*, *119*, 13,751–13,770. https://doi.org/10.1002/2014JD022113

Castle, S. L., Reager, J. T., Thomas, B. F., Purdy, A. J., Lo, M.-H., Famiglietti, J. S., & Tang, Q. (2016). Remote detection of water management impacts on evapotranspiration in the Colorado River Basin. *Geophysical Research Letters*, *43*, 5089–5097. https://doi.org/10.1002/2016GL068675

Cheng, L., Xu, Z., Wang, D., & Cai, X. (2011). Assessing interannual variability of evapotranspiration at the catchment scale using satellite-based evapotranspiration data sets. *Water Resources Research*, *47*, W09509. https://doi.org/10.1029/2011WR010636

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015a). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*, 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015b). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, *51*, 2515–2542. https://doi.org/10.1002/2015WR017200

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*, 2350–2365. https://doi.org/10.1002/2015WR017910

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*, W00B02. https://doi.org/10.1029/2007WR006735

Famiglietti, J., Lo, M., Ho, S., Bethune, J., Anderson, K., Syed, T., et al. (2011). Satellites measure recent rates of groundwater depletion in California's Central Valley. *Geophysical Research Letters*, *38*, L03403. https://doi.org/10.1029/2010GL046442

Gao, H., Tang, Q., Ferguson, C. R., Wood, E. F., & Lettenmaier, D. P. (2010). Estimating the water budget of major US river basins via remote sensing. *International Journal of Remote Sensing*, *31*(14), 3955–3978. https://doi.org/10.1080/01431161.2010.483488

Gao, H., Tang, Q., Shi, X., Zhu, C., Bohn, T., Su, F., et al. (2010). Water budget record from Variable Infiltration Capacity (VIC)model. In *Algorithm theoretical basis document* (Tech. Rep., pp. 120–173). Seattle, WA: Department of Civil Engineering, University of Washington.

Gulden, L. E., Rosero, E., Yang, Z.-L., Rodell, M., Jackson, C. S., Niu, G.-Y., et al. (2007). Improving land-surface model hydrology: Is an explicit aquifer model better than a deeper soil profile? *Geophysical Research Letters*, *34*, L09402. https://doi.org/10.1029/2007GL029804

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Haacker, E. M. K., Kendall, A. D., & Hyndman, D. W. (2015). Water level declines in the High Plains aquifer: Predevelopment to resource senescence. *Groundwater*, *54*(2), 231–242.

Hejazi, M. I., & Cai, X. (2009). Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. *Advances in Water Resources*, *32*(4), 582–593. https://doi.org/10.1016/j.advwatres.2009.01.009

Housh, M., Cai, X., Ng, T. L., McIsaac, G. F., Ouyang, Y., Khanna, M., et al. (2014). System of systems model for analysis of biofuel development. *Journal of Infrastructure Systems*, *21*(3), 04014050.

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

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*(7318), 951–954. https://doi.org/10.1038/nature09396

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362

Konikow, L. F. (2015). Long-term groundwater depletion in the United States. *Groundwater*, *53*(1), 2–9. https://doi.org/10.1111/gwat.12306

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., et al. (2004). Regions of strong coupling between soil moisture and precipitation. *Science*, *305*(5687), 1138–1140. https://doi.org/10.1126/science.1100217

Koster, R. D., Fekete, B. M., Huffman, G. J., & Stackhouse, P. W. (2006). Revisiting a hydrological analysis framework with International Satellite Land Surface Climatology Project Initiative 2 rainfall, net radiation, and runoff fields. *Journal of Geophysical Research*, *111*, D22S05. https://doi.org/10.1029/2006JD007182

Koster, R. D., & Suarez, M. J. (1999). A simple framework for examining the interannual variability of land surface moisture fluxes. *Journal of Climate*, *12*(7), 1911–1917.

Kumar, P. (2015). Hydrocomplexity: Addressing water security and emergent environmental risks. *Water Resources Research*, *51*, 5827–5838. https://doi.org/10.1002/2015WR017342

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*, W04531. https://doi.org/10.1029/2011WR011453

Lettenmaier, D. P., & Famiglietti, J. S. (2006). Hydrology: Water from on high. *Nature*, *444*(7119), 562–563. https://doi.org/10.1038/444562a

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*(D7), 14,415–14,428. https://doi.org/10.1029/94JD00483

Long, D., Longuevergne, L., & Scanlon, B. R. (2015). Global analysis of approaches for deriving total water storage changes from GRACE satellites. *Water Resources Research*, *51*, 2574–2594. https://doi.org/10.1002/2014WR016853

Mahrt, L., & Ek, M. (1984). The influence of atmospheric stability on potential evaporation. *Journal of Climate and Applied Meteorology*, *23*(2), 222–234. https://doi.org/10.1175/1520-0450(1984)023<0222:TIOASO>2.0.CO;2

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernandez-Prieto, D., et al. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*(5), 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017

McGuire, V. (2012). Water-level and storage changes in the High Plains aquifer, predevelopment to 2011 and 2009–11. In *U.S. Geological Survey Scientific Investigations Report*, *2012–5291* (15 pp.). Retrieved from http://pubs.usgs.gov/sir/2012/5291/

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*, D07S90. https://doi.org/10.1029/2003JD003823

Montanari, A., & Di Baldassarre, G. (2013). Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. *Advances in Water Resources*, *51*, 498–504. https://doi.org/10.1016/j.advwatres.2012.09.007

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., et al. (2013). "Panta Rhei—Everything flows": Change in hydrology and society—The IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Mu, Q., Zhao, M., Heinsch, F. A., Liu, M., Tian, H., & Running, S. W. (2007). Evaluating water stress controls on primary production in biogeochemical and remote sensing based models. *Journal of Geophysical Research*, *112*, G01012. https://doi.org/10.1029/2006JG000179

Mu, Q., Zhao, M., Kimball, J. S., McDowell, N. G., & Running, S. W. (2013). A remotely sensed global terrestrial drought severity index. *Bulletin of the American Meteorological Society*, *94*(1), 83–98. https://doi.org/10.1175/BAMS-D-11-00213.1

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

Munier, S., Polebistki, A., Brown, C., Belaud, G., & Lettenmaier, D. P. (2015). SWOT data assimilation for operational reservoir management on the upper Niger River Basin. *Water Resources Research*, *51*, 554–575. https://doi.org/10.1002/2014WR016157

Mutiibwa, D., & Irmak, S. (2013). AVHRR-NDVI-based crop coefficients for analyzing long-term trends in evapotranspiration in relation to changing climate in the U.S. High Plains. *Water Resources Research*, *49*, 231–244. https://doi.org/10.1029/2012WR012591

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

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2011). Multisource estimation of long-term terrestrial water budget for major Global River Basins. *Journal of Climate*, *25*(9), 3191–3206.

Pan, M., & Wood, E. F. (2006). Data assimilation for estimating the terrestrial water budget using a constrained ensemble Kalman filter. *Journal of Hydrometeorology*, *7*(3), 534–547. https://doi.org/10.1175/JHM495.1

Porporato, A., Feng, X., Manzoni, S., Mau, Y., Parolari, A. J., & Vico, G. (2015). Ecohydrological modeling in agroecosystems: Examples and challenges. *Water Resources Research*, *51*, 5081–5099. https://doi.org/10.1002/2015WR017289

Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., et al. (2015). The observed state of the water cycle in the early twenty-first century. *Journal of Climate*, *28*(21), 8289–8318. https://doi.org/10.1175/JCLI-D-14-00555.1

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C. J., et al. (2004). The global land data assimilation system. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., & McCabe, M. F. (2009). Closing the terrestrial water budget from satellite remote sensing. *Geophysical Research Letters*, *36*, L07403. https://doi.org/10.1029/2009GL037338

Shuttleworth, W., Gurney, R., Hsu, A., & Ormsby, J. (1989). FIFE: The variation in energy partition at surface flux sites. *IAHS Publication*, *186*, 67–74.

Shuttleworth, W. J. (2007). Putting the "vap" into evaporation. *Hydrology and Earth System Sciences*, *11*(1), 210–244. https://doi.org/10.5194/hess-11-210-2007

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

Strassberg, G., Scanlon, B. R., & Chambers, D. (2009). Evaluation of groundwater storage monitoring with the GRACE satellite: Case study of the High Plains aquifer, central United States. *Water Resources Research*, *45*, W05410. https://doi.org/10.1029/2008WR006892

Swenson, S. C., & Lawrence, D. M. (2015). A GRACE-based assessment of interannual groundwater dynamics in the Community Land Model. *Water Resources Research*, *51*, 8817–8833. https://doi.org/10.1002/2015WR017582

Tang, Q., Peterson, S., Cuenca, R. H., Hagimoto, Y., & Lettenmaier, D. P. (2009). Satellite-based near-real-time estimation of irrigated crop water consumption. *Journal of Geophysical Research*, *114*, D05114. https://doi.org/10.1029/2008JD010854

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

Vogel, R. M., Lall, U., Cai, X., Rajagopalan, B., Weiskel, P. K., Hooper, R. P., & Matalas, N. C. (2015). Hydrology: The interdisciplinary science of water. *Water Resources Research*, *51*, 4409–4430. https://doi.org/10.1002/2015WR017049

Wang, D., & Cai, X. (2009). Detecting human interferences to low flows through base flow recession analysis. *Water Resources Research*, *45*, W07426. https://doi.org/10.1029/2009WR007819

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Brewer, M., et al. (2016). Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research: Atmospheres*, *121*, 2750–2779. https://doi.org/10.1002/2015JD023733

Xia, Y., Hobbins, M. T., Mu, Q., & Ek, M. B. (2015). Evaluation of NLDAS-2 evapotranspiration against tower flux site observations. *Hydrological Processes*, *29*(7), 1757–1771. https://doi.org/10.1002/hyp.10299

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., et al. (2012). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *Journal of Geophysical Research*, *117*, D03110. https://doi.org/10.1029/2011JD016051

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*, D03109. https://doi.org/10.1029/2011JD016048

Xia, Y., Mocko, D., Huang, M., Li, B., Rodell, M., Mitchell, K. E., et al. (2017). Comparison and assessment of three advanced land surface models in simulating terrestrial water storage components over the United States. *Journal of Hydrometeorology*, *18*(3), 625–649. https://doi.org/10.1175/JHM-D-16-0112.1

Xia, Y., Peter-Lidard, C. D., Huang, M., Wei, H., & Ek, M. (2015). Improved NLDAS-2 Noah-simulated hydrometeorological products with an interim run. *Hydrological Processes*, *29*(5), 780–792. https://doi.org/10.1002/hyp.10190

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

Zeng, R., & Cai, X. (2014). Analyzing streamflow changes: Irrigation-enhanced interaction between aquifer and streamflow in the Republican River basin. *Hydrology and Earth System Sciences*, *18*(2), 493–502. https://doi.org/10.5194/hess-18-493-2014

Zeng, R., & Cai, X. (2015). Assessing the temporal variance of evapotranspiration considering climate and catchment storage factors. *Advances in Water Resources*, *79*, 51–60. https://doi.org/10.1016/j.advwatres.2015.02.008

Zeng, R., & Cai, X. (2016). Climatic and terrestrial storage control on evapotranspiration temporal variability: Analysis of river basins around the world. *Geophysical Research Letters*, *43*, 185–195. https://doi.org/10.1002/2015GL066470

Zhang, K., Kimball, J. S., Nemani, R. R., & Running, S. W. (2010). A continuous satellite-derived global record of land surface evapotranspiration from 1983 to 2006. *Water Resources Research*, *46*, W09522. https://doi.org/10.1029/2009WR008800