https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Evaluating the Statistical Agreement between Gridded Evapotranspiration Data Sets in the Conterminous United States via Triple Collocation

Keith Doore[1], Thomas M. Over[2], Timothy O. Hodson[3], and Sydney S. Foks[4]

[1]U.S. Geological Survey Central Midwest Water Science Center, Iowa City, IA, USA
[2]U.S. Geological Survey Central Midwest Water Science Center, Urbana, IL, USA
[3]U.S. Geological Survey Water Resources Mission Area, Urbana, IL, USA
[4]U.S. Geological Survey Water Resources Mission Area, Tacoma, WA, USA

Correspondence: Keith Doore (kdoore@usgs.gov; present address: kjdoore@gmail.com)

Abstract.

Evapotranspiration (ET) is a critical component in hydrologic budgets and one of the most difficult to measure, which can create substantial inconsistencies among ET datasets. This is particularly an issue for benchmarking the performance of hydrologic model simulations, as we lack any ground truth gridded data for verifying ET datasets. Commonly, some form of collocation analysis is used to estimate the error variance of ET data from multiple independent datasets. However, this technique only assesses the error variances and does not estimate the biases of the datasets from the truth, which can be more important for some applications. Although the biases from the truth cannot be determined, relative biases between datasets can. To assess these, we developed a novel method that combines the temporal median of the relative bias between datasets and the error covariance matrices estimated from the Extended Collocation method to derive dataset agreement probabilities. We then applied this method to six gridded monthly ET datasets that cover the Conterminous United States (CONUS): SSEBop, GLEAM, ERA5-Land, NLDAS-2, TerraClimate, and the water balance ET dataset from Reitz et al. (2023a). From these probabilities, we found that all but one dataset pair had >70% of grid cells with p-values > 0.16 (and >85% of grid cells with p-values > 0.05) across CONUS over the full time period, which indicates reasonable agreement. When split by season, winter has 36.8% of grid cells across all dataset pairs with p-values > 0.16 (56.7% with p-values > 0.05), with spring, summer, and fall having 41.9% (63.3%), 54.2% (74.4%), and 63.3% (82.2%), respectively. However, when looking at certain regions of concern for water resources management, estimated agreement probabilities split by season showed that the Central Valley had a majority of dataset pairs with p-values < 0.005 in the summer. This may be directly attributed to the lack of an irrigation component in the GLEAM, ERA5-Land, NLDAS-2, and TerraClimate datasets.

1 Introduction

Hydrological modeling is a powerful tool for studying continental-scale hydrologic processes that are critical to water management such as predicting river flow and quantifying changes in groundwater storage (Beven, 2012; Archfield et al., 2015; Siqueira et al., 2018; Zhu et al., 2023). Realistic model simulations at this scale require accurate, long-term, and spatiotem-

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

porally consistent data sets of major hydrologic components such as precipitation, soil moisture, and evapotranspiration (Saxe et al., 2021; Singh et al., 2024). Direct in situ observational data sets of these components are typically discontinuous in space and time and can have significant uncertainty and biases between measurement methods (Di Baldassarre and Montanari, 2009; Zhou et al., 2021; Levin et al., 2023). Alternatively, ex situ gridded data sets estimated from various methods can have uncertainty and biases due to sensor calibration and instrumentation errors (Chen et al., 2014; Bhatt et al., 2018), insufficient model parametrization (Mallick et al., 2018; Trebs et al., 2021), sub-grid heterogeneity (He et al., 2021; Torres-Rojas et al., 2022; Waterman et al., 2024), and/or lack of sufficient validation measurements (Gruber et al., 2020). The uncertainties within and inconsistencies between component data sets are then propagated through the hydrological models leading to less accurate simulations. Consequently, quantifying the uncertainty and using it to evaluate the statistical agreement between these component data can lead to a better understanding of their shortcomings, thereby facilitating improvements to the component data and subsequently improvements in hydrological modeling accuracy and consistency (McMillan et al., 2018).

Of these hydrologic components, actual evapotranspiration (ET), the amount of water that changes phase from liquid to vapor and is returned to the atmosphere from all surfaces and plant transpiration, is the second largest component of the water cycle behind precipitation and is the primary driver that connects the energy, water, and carbon cycles (Arnold et al., 1999; Oki and Kanae, 2006; Alexandris, 2013; Fisher et al., 2017; Migliavacca et al., 2021; Guse et al., 2021). While being one of the most critical components, ET is also one of the most uncertain and inconsistent with total annual estimates for the conterminous United States (CONUS) differing by up to $300\ \mathrm{mm\,yr^{-1}}$ or $\approx 75\%$ (Saxe et al., 2021; Li et al., 2022). ET can be estimated with a variety of methods including in situ measurements, remote-sensing, reanalysis, and interpolation of in situ measurements (Levin et al., 2023). In situ measurements typically consist of temporally sporadic and spatially fixed estimates from simple pan evaporation, eddy-covariance flux towers, or lysimeter systems, which measure ET at the local ($< 1\ \mathrm{km^2}$) scale (e.g., Yang et al., 2012; Pastorello et al., 2020). Alternatively, remote-sensing ET estimates are generated at temporally continuous and continental/global scales by utilizing energy and/or water balance models with physical, empirical, or semi-empirical methods that calculate ET from directly observable satellite-based and/or meteorological data products (e.g., Anderson et al., 2007; Mu et al., 2011; Senay et al., 2013; Martens et al., 2017; Zhang et al., 2019; Fisher et al., 2020). As for reanalysis ET products, they utilize long-term ground-based meteorological data supplemented by satellite-based observations to produce high temporal and spatial resolution hydrologic data sets commonly spanning back to the mid 1900s (e.g., Rodell et al., 2004; Bosilovich et al., 2008; Reichle et al., 2017; Muñoz-Sabater et al., 2021). Finally, interpolation ET products interpolate in situ measurements or other local ET estimates to a continuous spatiotemporal grid using hydrologic and climatic variables as guiding variables (e.g., Reitz et al., 2017; Abatzoglou et al., 2018; Jung et al., 2019; Reitz et al., 2023a).

Each method for estimating ET has its own sources of uncertainty. For example, in situ measurements can be highly accurate at local scales, but they are limited to this local scale due to ET being highly dependent on the land cover, topography, and other environmental conditions associated with the instrument location (Friedl et al., 2002; Richardson et al., 2006; Mu et al., 2007; Choi et al., 2009; Zheng et al., 2017; Tang and Li, 2017). Remote-sensing and reanalysis estimates, while applicable to broader scales, can inherit uncertainty and biases from their source satellite and meteorological data products as well as assumptions built into methods and models used to calculate ET from these input data. As for interpolation products, while they may start

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

with higher accuracy ET measurements, interpolation methods simplify spatial heterogeneity and other local complexities and thereby introduce uncertainty and potential biases into the final product. Therefore, quantifying the complex uncertainty and relative biases of these various ET products can give insight on how individual methods can be improved based on differences between the estimation methods.

Various methods exist for estimating the uncertainty within large-scale data sets, but most of these methods require a high-quality reference data set whose uncertainty is known. This can be a steep requirement that is impossible to meet for ET and other types of climate data when performing analyses at continental to global scales. For example, Volk et al. (2023) demonstrated how complex data processing, bias correction, and modeling efforts can be in order to turn reference flux tower ET observations into data that can be used as a reference for gridded ET. Even then, in situ data like this can only be used as a reference for a small fraction of grid cells in a gridded data set. As a result, statistical collocation analysis methods have been developed which can estimate the uncertainty within data sets without the need of assuming one as a reference. In its basic form, referred to as Triple Collocation (TC), a collocation analysis requires three spatiotemporally collocated data sets, from which it estimates the random error variances for each collocated data set (Stoffelen, 1998). Beyond TC, collocation analysis methods have been expanded and modified to allow for estimating error covariances (Zwieback et al., 2012; Gruber et al., 2016a), reducing the requirement to have three unique data sets by including a time lagged data set (Su et al., 2014; Dong et al., 2019, 2020b), and calculating the correlation of each data set with the unknown truth (McColl et al., 2014). The collocation analysis method has been extensively applied to a variety of variables including ocean wind and wave height products (Stoffelen, 1998; Caires and Sterl, 2003; Montoya and Osorio, 2014; Ribal and Young, 2020), soil moisture (Scipal et al., 2008; Miralles et al., 2010; Yilmaz et al., 2012; Su et al., 2014; Gruber et al., 2017; Chen et al., 2018; Kim et al., 2018; Dong et al., 2020a; Xu et al., 2021), precipitation (Roebeling et al., 2012; Alemohammad et al., 2015; Massari et al., 2017; Li et al., 2018; Tang et al., 2020; Duan et al., 2021; Lyu et al., 2021), leaf area index (Fang et al., 2012; Jiang et al., 2017), and total water storage (Van Dijk et al., 2014; Yin and Park, 2021; Baik et al., 2022).

As for ET, most studies implement collocation analyses in one of two ways. The first uses the method to derive uncertainties for various ET data products to better understand their advantages and limitations across different regions and land cover types (Khan et al., 2020; Ochege et al., 2021; Li et al., 2022; Jia et al., 2022). These applications commonly find that the ET data sets have different performance levels (i.e., varying levels of uncertainty) across these land cover types, which is a result of the method used to derive the ET product (Khan et al., 2020). The other implementations try to minimize these variations in performance by merging the ET data sets into a single data set using a weighting determined by the collocation uncertainty (Khan et al., 2018; Park et al., 2023; Li et al., 2023).

In this study, we present a novel third type of application of collocation analysis that uses the method's estimated error variances and covariances to evaluate the statistical agreement between ET data products. We then apply it to six gridded CONUS-wide ET data sets and examine in detail the results for three regions in CONUS: the Central Valley, High Plains aquifer region, and Upper Colorado River Basin. These regions are highlighted in this study because they are of increased concern to water resource management, as they are heavily agricultural and recently have been experiencing extreme drought and accelerated declines in aquifer storage (Faunt, 2009; Stanton et al., 2011; Tillman et al., 2022). By having an assessment

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

of the agreement between ET data products and their uncertainty, it could help improve water availability assessments within these regions.

The paper is structured as follows: in Section 2 we describe the collocation analysis methodology in detail, along with our novel method for evaluating the statistical agreement between data sets. We then present the six CONUS-wide ET data sets in Section 3, along with details of the three regions over which we spatially aggregate the ET data sets. In Section 4.1, we apply the collocation analysis method to the ET data sets to estimate their error variances and covariances. With the collocation results, we apply our novel method for evaluating the agreement between data sets in Section 4.2. We then discuss the findings of the agreement evaluation in Section 5. Finally, we summarize our results in Section 6. All workflows used in this paper are available as Jupyter Notebooks, which are linked in the *Code availability* section.

2 Methods

2.1 Collocation Analysis Methodology

Collocation analysis is a method for approximating the error variance of a stochastic property measured by multiple collocated measurement/observing systems[1]. This error variance can be estimated without requiring any of the measurement systems to have observed the "true" value. Typically, the collocation analysis method makes four assumptions about the measurement systems that are critical to its validity (Zwieback et al., 2012; Gruber et al., 2016b):

1. the observed signal and random errors are stationary (i.e., the mean and variance of each is constant with time),
2. the errors have no cross-correlation (i.e., measurement system errors are independent of each other),
3. each measurement system has orthogonal errors (i.e., the measurement system errors are independent of the true value), and
4. the errors have no autocorrelation (i.e., the errors are not correlated with time).

In practice, it cannot always be assured that all of these assumptions are met, especially assumptions (2) and (3) (Yilmaz and Crow, 2014). Several studies have proposed expanded collocation methods to account for breaking these assumptions (Su et al., 2014; Pierdicca et al., 2017; Nearing et al., 2017; Dong et al., 2019). However, these expanded methods typically impose other alternative assumptions. In order to not impose alternative assumptions, we used the generalized form of the TC method, known as Extended Collocation (EC; Zwieback et al., 2012; Gruber et al., 2016a), which keeps the TC assumptions but accommodates an arbitrary number of measurement systems rather than just three.

[1]We use the term "measurement system" in Section 2 for consistency with the collocation literature. It is a general term that refers to any data set (e.g., in situ, remotely sensed, etc.). They can be considered collocated in terms of both exact collocation (i.e., aligned gridded data) and relative collocation (i.e., gridded data covering an in situ site).

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

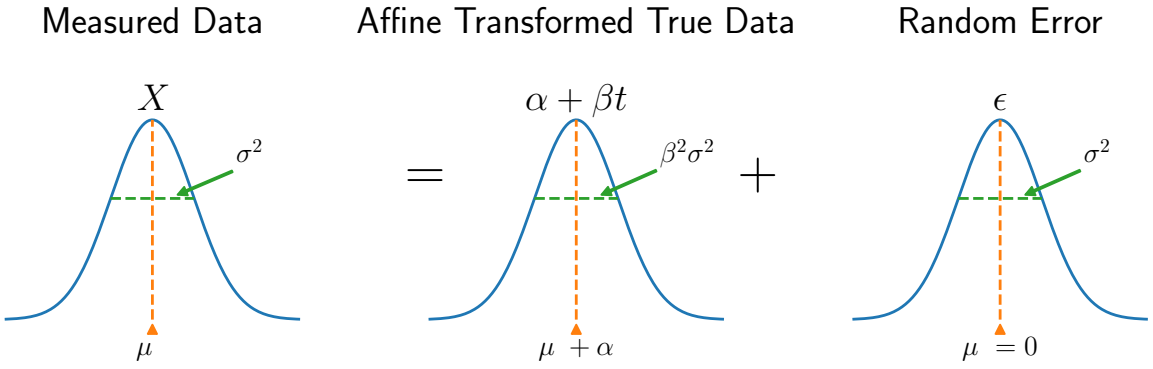

Figure 1. A visual representation of the affine error model assumed by our implementation of Triple Collocation (TC). The blue bell curves represent the distributions, from left to right, of the measured data ($\boldsymbol{X}$), the affine transformed true data ($\alpha + \beta\boldsymbol{t}$, where α and β are the constant affine model parameters at a given location and $\boldsymbol{t}$ is the measurements' true values), and the additive random errors ($\boldsymbol{\varepsilon}$). The population means for each corresponding distribution are shown below the vertical orange lines (μ_X , $\mu_t + \alpha$, $\mu_\varepsilon = 0$, respectively), and the variances of the distributions are shown next to the horizontal green lines (σ_X^2, $\beta^2\sigma_t^2$, σ_ε^2, respectively).

2.1.1 Triple Collocation

The basic formulation of TC commonly assumes an affine error model (which imposes an additional assumption of linearity) relating the observations to the true value and error (Zwieback et al., 2012; McColl et al., 2014):

$$\boldsymbol{X}_i = \alpha_i + \beta_i \boldsymbol{t} + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{X}_i$ is the measured time series at a given location from the ith collocated measurement system, α_i and β_i are the constant affine model parameters at the given location, $\boldsymbol{t}$ is the measurements' true values, and $\boldsymbol{\varepsilon}_i$ is the additive random errors. A visual representation of the affine error model is shown in Figure 1.

There are two general methods/notations for approximating the error variances (i.e., $\sigma^2_{\varepsilon_i}$): the *difference* notation and the *covariance* notation. Although the two notations both approximate $\sigma^2_{\varepsilon_i}$, the distinction between the two is that the difference notation rescales the measurement systems to the data space of a system arbitrarily chosen as a reference. It then requires the derived error variances of the unscaled measurement systems to be rescaled back to their native data space, whereas the covariance notation does not. So, we utilized the covariance notation instead of the difference notation to avoid any rescaling steps.

Solving for the covariance of two measurement systems (i and j) using Equation 1 gives

$$\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \beta_i\beta_j\sigma^2_{\boldsymbol{t}} + \beta_i\mathrm{Cov}(\boldsymbol{t}, \boldsymbol{\varepsilon}_j) + \beta_j\mathrm{Cov}(\boldsymbol{t}, \boldsymbol{\varepsilon}_i) + \mathrm{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j), \tag{2}$$

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

where $\sigma_{\boldsymbol{t}}^2 = \mathrm{Var}(\boldsymbol{t})$. This equation can then be simplified by incorporating the TC assumptions above to give

$$\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} \beta_i \beta_j \sigma_{\boldsymbol{t}}^2, & \text{for } i \neq j \\ \beta_i^2 \sigma_{\boldsymbol{t}}^2 + \sigma_{\varepsilon_i}^2, & \text{for } i = j \end{cases}, \tag{3}$$

since assumption (2) gives $\mathrm{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = 0$ for $i \neq j$, and assumption (3) gives $\mathrm{Cov}(\boldsymbol{t}, \boldsymbol{\varepsilon}_i) = 0$. Solving the system of equations in Equation 3 is only possible with three or more measurement systems (to have at least the same number of equations as unknowns) and results in the following estimates of the error variances[2]:

$$\sigma_{\varepsilon_i}^2 = \sigma_i^2 - \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}}, \tag{4}$$

where $\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sigma_{ij}$ (i.e., the covariance between the i and j measurement systems' spatiotemporally collocated time series at a given location), and k is the third measurement system. Therefore, an estimate of the error variance of a measurement system at a collocation can be calculated simply from the difference between its time series variance and a ratio of the time series covariance between the collocated system triplet.

2.1.2 Extended Collocation

An extended generalized form of the collocation technique was originally formulated in Zwieback et al. (2012) to allow for the use of three or more measurement systems. By adding additional collocated measurement systems, it was shown that some cross-correlations between errors could be allowed and potentially estimated. More specifically, EC generalizes the contents of TC assumption (2) to be:

2. the errors may have cross-correlation between measurement systems if each correlated measurement system is also a member of at least one measurement system triplet that has no cross-correlation of errors.

This means EC still requires three independent measurement systems like TC, but by including additional systems, error covariances can be estimated for certain correlated pairs in addition to the error variances. However, one caveat is that the systems whose errors are cross-correlated (i.e., are covariant) must be known a priori.

With the update to assumption (2), the simplification of Equation 2 changes to

$$\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} \beta_i^2 \sigma_{\boldsymbol{t}}^2 + \sigma_{\varepsilon_i}^2, & \text{for } i = j \\ \beta_i \beta_j \sigma_{\boldsymbol{t}}^2, & \text{for } i \neq j \text{ where } \sigma_{\varepsilon_i,\varepsilon_j} = 0 \\ \beta_i \beta_j \sigma_{\boldsymbol{t}}^2 + \sigma_{\varepsilon_i,\varepsilon_j}, & \text{for } i \neq j \text{ where } \sigma_{\varepsilon_i,\varepsilon_j} \neq 0 \end{cases}, \tag{5}$$

where $\sigma_{\varepsilon_i,\varepsilon_j}$ is the error covariance. Solving this system of equations when $\sigma_{\varepsilon_i,\varepsilon_j} \neq 0$ is only possible with more than three observing systems, three of which must have $\sigma_{\varepsilon_i,\varepsilon_j} = 0$ to satisfy assumption (2). With the measurement system quadruplet i,

[2]Solving this system of equations requires treating $\beta_i \sigma_{\boldsymbol{t}}$ as a single unknown to reduce the number of unknowns by 1. Additionally, β_i, α_i, and $\sigma_{\boldsymbol{t}}$ cannot be directly calculated using the covariance notation of TC.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

j, k, and l (assuming only $\sigma_{\varepsilon_i,\varepsilon_j} \neq 0$), the error variances can be estimated via Equation 4. The error covariance of i and j can be estimated as

$$\sigma_{\varepsilon_i,\varepsilon_j} = \sigma_{ij} - \frac{\sigma_{ik}\sigma_{jl}}{\sigma_{kl}}. \tag{6}$$

With the covariance, cross-correlation of errors can easily be estimated by combining the results of Equations 4 and 6 as

$$\rho_{\varepsilon_i,\varepsilon_j} = \frac{\sigma_{\varepsilon_i,\varepsilon_j}}{\sqrt{\sigma^2_{\varepsilon_i}\sigma^2_{\varepsilon_j}}}, \tag{7}$$

where $\rho_{\varepsilon_i,\varepsilon_j}$ is the error cross-correlation.

While Equations 4, 6, and 7 are shown for measurement system triplets or quadruplets, they can easily be applied to additional measurement systems beyond the required minimum. One thing to note is that as the number of measurement systems in the EC evaluation increases, so do the possible combinations of systems. If more than three systems are considered independent, then these different possible combinations can result in multiple estimates of the error variance and covariances for each system. In practice, these estimates for a given system may not result in exactly the same values, since random noise in the data can cause variations in data covariances. However, they should be reasonably close, assuming the measurement systems truly are independent of each other. Thus, substantial deviations in error variances and covariances from multiple estimates for a given system are an indicator of the presence of cross-correlated errors within the assumed independent triplet (i.e., the measurement systems are not truly independent of each other).

2.2 Statistical Agreement Probabilities

Collocation analysis alone only provides error variances and covariances and does not provide any information on the mean differences (relative biases) between measurement systems. Such differences are easy to compute, but whether the differences are statistically significant depends on their error. For example, when looking at ET, does a bias of $300\ \mathrm{mm\,yr^{-1}}$ indicate that the measurement systems are truly biased if the bias has an error of $200\ \mathrm{mm\,yr^{-1}}$? To address this question, we developed a simple analytical method that converts the EC error covariance matrices (i.e., the results from Equations 4 and 6 grouped into an error covariance matrix) into a bias error, which can then be used to evaluate whether the measured bias between measurement systems is statistically significant. To estimate the statistical significance, the method combines the relative bias and bias error into what we term "agreement probabilities", which is a consistent way of quantifying significance between all measurement system pairs that may have highly different biases and bias errors.

To calculate agreement probabilities, the absolute relative biases between each measurement system pair are first derived from:

$$\boldsymbol{B}_{ij} = |\boldsymbol{X}_i - \boldsymbol{X}_j|, \tag{8}$$

where $\boldsymbol{B}_{ij}$ is the element-wise absolute relative bias (i.e., the bias between each time series element). The absolute relative bias is utilized because for agreement it is not important which measurement system has larger values, rather the magnitude of

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

the bias is what is important. Next, the error variances and covariances in the EC error covariance matrices are propagated to a bias error $\sigma_{\varepsilon_{\boldsymbol{B}ij}}$:

$$\sigma_{\varepsilon_{\boldsymbol{B}ij}} = \sqrt{\sigma^2_{\varepsilon_i} + \sigma^2_{\varepsilon_j} - 2\sigma_{\varepsilon_i,\varepsilon_j}}. \tag{9}$$

With the bias error, a "bias distribution" is made from a zero-mean normal distribution with a standard deviation of the bias error. Finally, using this bias distribution, the agreement probability between the measurement system pair is calculated and defined as the density of the distribution that is greater than or equal to the median absolute relative bias. The median absolute relative bias (i.e., the median bias of the time series elements) is utilized to have a time-averaged value like the EC error covariance matrices and to make the results more tractable (i.e., a single agreement probability versus one for each bias time series element). We note that the choice of using the median is relatively arbitrary and that other approaches could be reasonably taken at this step (e.g., the mean).

To clarify the meaning of the agreement probability, it is the probability that a bias value greater than or equal to the median absolute bias between the measurement system pair is purely due to random error in the data, given the estimated errors of the pair from EC. The value of the probability ranges from 0 to 0.5 (as the absolute relative bias cannot be less than 0), with values near 0 indicating that there is little to no probability the bias is due the errors (i.e., the data are truly biased) and values near 0.5 indicating that there is no distinguishable bias between the measurement systems (i.e., $\boldsymbol{B}_{ij} \approx 0$). As the agreement probability is equivalent to a one-tailed test, it can also be thought of as the p-value associated with a statistical hypothesis test with the null hypothesis that the measurement system pair are not biased (i.e., the measurement systems agree to the extent that can be determined given their errors). Any significance level can then be chosen to determine if the measurement systems are not in agreement. Besides the ability to perform a statistical hypothesis test, the primary advantage of the agreement probability over quoting the median bias and error (e.g., $300 \pm 200\ \mathrm{mm\,yr^{-1}}$) is that it is scale invariant, meaning a bias and error of $300 \pm 200\ \mathrm{mm\,yr^{-1}}$ will result in the same agreement probability as $150 \pm 100\ \mathrm{mm\,yr^{-1}}$. This feature is highly useful when comparing multiple measurement systems as it ensures that measurement systems with highly variable scales can be fairly compared.

3 Evapotranspiration Data

This study evaluates six commonly used ET products with full CONUS coverage: SSEBop (Senay et al., 2013; Senay, 2018), GLEAM v3.7b (Miralles et al., 2011; Martens et al., 2017), ERA5-Land (Muñoz-Sabater et al., 2021), NLDAS-2 (Noah) (Xia et al., 2012), TerraClimate (Abatzoglou et al., 2018), and a streamgage-based water balance product (WBET; Reitz et al., 2023a). As ET products are typically calculated from other remotely sensed reference data, we selected these six data sets as they span a variety of reference data types and calculations methods. By including a variety of reference data types and calculation methods, it is commonly assumed that any cross-correlation of errors between the data sets should be minimized and satisfy assumption (2) of EC at a basic level (e.g., Scipal et al., 2008; Gruber et al., 2016b; Jia et al., 2022). We discuss this assumption in detail in Section 5.3. Finally, we selected these products as they had versions that were already aggregated

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Table 1. Characteristics of the evapotranspiration data sets used in this study.

Data Sets	Reference Data Type	Calculation Method	Date Range	Resolution	Units	Data Reference
SSEBop	Ex situ	Energy balance	2001/01–2022/12	1 km	mm month^{-1}	Senay and Kagone (2019)
GLEAM v3.7b	Ex situ	Energy balance	2003/01–2022/12	0.25°	mm month^{-1}	Martens et al. (2017)
ERA5-Land	Reanalysis	Energy balance	1950/01–2022/12	0.1°	m day^{-1}	Muñoz Sabater, J. (2019)
NLDAS-2 (Noah)	Reanalysis	Energy balance	1979/02–2022/12	0.125°	kg m^{-2} month^{-1}	NLDAS Project (2020)
TerraClimate	Interpolated	Water balance	1958/01–2022/12	0.0416°	mm month^{-1}	Abatzoglou et al. (2017)
WBET	Interpolated	Water balance	1895/10–2018/09	800 m	mm day^{-1}	Reitz et al. (2023b)

(or could be aggregated at the data source) to a monthly time step and have temporal coverage through the vast majority of the 21st century. A summary of the ET data sets and their attributes is shown in Table 1.

3.1 SSEBop

The Operational Simplified Surface Energy Balance (SSEBop) model estimates an ET fraction using an energy balance method that incorporates land surface and maximum air temperature to solve for latent heat flux via a satellite psychrometric approach (Senay et al., 2013; Senay, 2018). This ET fraction is then used to convert a reference (i.e., potential) ET to actual ET. For the SSEBop data used in our study[3], land surface temperatures from the Moderate Resolution Imaging Spectroradiometer (MODIS; data product MOD11A2 version 6) were utilized along with air temperature data from the Parameter-elevation Regressions on Independent Slopes Model (PRISM; http://www.prism.oregonstate.edu/) to generate the ET fraction. The ET fraction data were then used to scale reference ET from gridMET (Abatzoglou, 2013) to generate the actual ET estimates.

3.2 GLEAM v3.7b

In this study, we used the Global Land Evaporation Amsterdam Model version 3.7b (GLEAM 3.7b; Miralles et al., 2011; Martens et al., 2017)[4]. The model dissects ET into its various components (bare-soil evaporation, open-water evaporation, sublimation, transpiration, and interception loss) and independently estimates each. Version b calculates ET utilizing only satellite-based observations. First, potential total evaporation (i.e., evaporation, sublimation, and transpiration) is calculated with the energy balance-based Priestley and Taylor (1972) equation with observations of surface net radiation and near-surface air temperature. Then, using the Gash analytical model, interception loss is calculated from precipitation data. With these components, actual total evaporation (i.e., actual ET) is then estimated for the various land regions based on satellite observations of Vegetation Optical Depth (VOD) and soil moisture. The details of the input satellite data can be found in Table 1 of Martens et al. (2017).

[3]Data retrieved from https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/uswem/web/conus/eta/modis_eta/monthly/downloads/
[4]Data retrieved from https://www.gleam.eu/.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

3.3 ERA5-Land

The updated land component of the fifth generation of European Reanalysis (ERA5-Land; hereon referred to as ERA5) is a global reanalysis data set generated by re-running the land component of the original ERA5 climate reanalysis at a higher spatial resolution (Muñoz-Sabater et al., 2021). This increased resolution is created by interpolating the original atmospheric forcing datasets from a resolution of 0.25° to 0.1°. Like the original reanalysis data set, ERA5-Land products are created using the Carbon Hydrology-Tiled ECMWF Scheme for Surface Exchanges over Land (CHTESSEL) as its land surface model. However, ERA5-Land uses the updated model cycle version Cy45r1 of CHTESSEL compared to version Cy41r2 of the original ERA5. This version has several updates important to ET, including the addition of climatological seasonality of vegetation and a new method for estimating bare soil evaporation (Muñoz-Sabater et al., 2021).

3.4 NLDAS-2 (Noah)

Phase 2 of the North American Land Data Assimilation System (NLDAS-2) is a land-surface model product derived from ground and satellite- based observations for central North America (Xia et al., 2012). NLDAS-2 runs multiple land surface models including Noah, Mosaic, and Variable Infiltration Capacity (VIC) to generate quality-controlled and consistent data sets of various land surface states and fluxes. The system was designed to minimize errors in soil moisture and energy stores, which often result in accuracy issues for numerical weather prediction models. For our study, we used the ET data generated from the Noah land surface model at the monthly temporal and 0.125° spatial resolution (i.e., NLDAS_NOAH0125_M). The Noah model of NLDAS was chosen due to its historical use with the Weather Research and Forecasting (WRF) models.

3.5 TerraClimate

TerraClimate is a climatically aided interpolation data set of monthly global climate and water balance surface variables (Abatzoglou et al., 2018)[5]. The climatically aided interpolation method combines a higher-temporal, lower-spatial resolution data set via bilinear interpolation of its temporal anomalies with a lower-temporal, higher-spatial resolution dataset to create a higher-temporal and -spatial resolution product (Willmott and Robeson, 1995). In terms of TerraClimate, the higher-temporal resolution data was taken from Climate Research Unit time series data version 4.0 (CRU Ts4.0; Harris et al., 2014) and the Japanese 55-year Reanalysis (JRA55; Kobayashi et al., 2015), while the higher-spatial resolution data utilized the WorldClim data set (Fick and Hijmans, 2017). To derive ET, the interpolated climate variables were passed through a modified Thornthwaite-Mather climatic water-balance model (Willmott et al., 1985; Dobrowski et al., 2013), which utilized the Wang-Erlandsson et al. (2016) extractable soil water storage capacity.

3.6 Water Balance ET

The water balance ET (WBET) data consisted of the interpolated monthly data presented in Reitz et al. (2023a). This method begins by deriving annual ET of gaged watersheds throughout CONUS using a water balance equation, which calculates ET

[5]Data retrieved from https://www.climatologylab.org/terraclimate.html

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

as the difference between the precipitation within the watershed, the stream discharge leaving the watershed, and any water storage change from groundwater-sourced irrigation. The data used for this calculation included PRISM precipitation data, USGS discharge data, and USGS county-level groundwater use data. The resulting ET from the gaged watersheds was then compared with ET estimated from a variety of ET equations (see Table 2 of Reitz et al., 2023a). The relative performance of this comparison was then converted into relative weights using a modified Bayesian model, average weighted ensemble approach. To have a CONUS-wide gridded product, the relative weights of each equation were extrapolated to a grid using random forest regression with a variety of explanatory variables (e.g., minimum and maximum temperature, wind speed, elevation, soil properties, etc.). The final ET estimates were then derived for the CONUS grid by calculating ET from each equation, combining the estimates with the extrapolated weight, and scaling the annual total ET to be equal to that derived from the water balance equation. However, before the final ET was calculated, an evaluation was performed to test the significance of the contribution of each equation to the ensemble. For the monthly data, it was found that ET estimated by the Fu-Zhang (Fu, 1981; Zhang et al., 2004) equation with Hamon (1961) potential ET resulted in the optimal ET estimate compared to the observed data. Therefore, the resulting monthly ET presented in Reitz et al. (2023a) is simply produced by calculating ET from the Fu-Zhang equation and scaling it to match the annual total ET derived from the water balance equation.

3.7 Data Processing

As seen in Table 1, the six data sets do not have the same spatial resolution nor do they have matching units. As this is a requirement to properly perform our analysis, we processed each data set to be collocated as well as have common units[6]. Due to it being the most common unit of our ET estimates, we performed a unit conversion to obtain all ET data in $\mathrm{mm\,month^{-1}}$, assuming a water density of $1\ \mathrm{g\,cm^{-3}}$. Since the data sets were selected to be aggregated by month, they are already temporally collocated within their overlapping data ranges. We did not further limit them to only span the overall common date range (i.e., 2003–2018). Instead, we waited until the selection of a quadruplet during the collocation analysis in Section 4 and then limited the dates to those common between each set. For example, the combination of SSEBop, GLEAM, ERA5, and WBET would result in a date range of 2003/01 to 2018/09, whereas the combination of ERA5, NLDAS, TerraClimate, and WBET would span 1979/01 to 2018/09. This choice is made to maximize the data usage, which helps minimize the variation in the estimated error variances (Zwieback et al., 2012).

To complete the collocation of the data sets, we needed to spatially collocate the data to a common grid. As discussed above, GLEAM, ERA5, and TerraClimate are global data sets, whereas SSEBop, NLDAS, and WBET are limited to the CONUS extent. To have a consistent spatial range, we first limited the data sets to the same latitude and longitude range (i.e., 24°N–53°N, 126°W–66°W). We then regridded the data to the lowest resolution data set (i.e., GLEAM) using a conservative regridding method (i.e., an area-weighted average). Additionally, missing data were propagated during regridding using a zero threshold policy; meaning if any source grid cell in the weighted average had missing data, then the target cell would be set as

[6]The ERA5 variable convention presents the ET data as negative values to indicate evaporation and positive values indicate condensation. To match the other data sets, we inverted the sign to have ET as positive values and set any remaining negative values (indicating condensation) to zero, prior to performing any other data processing.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

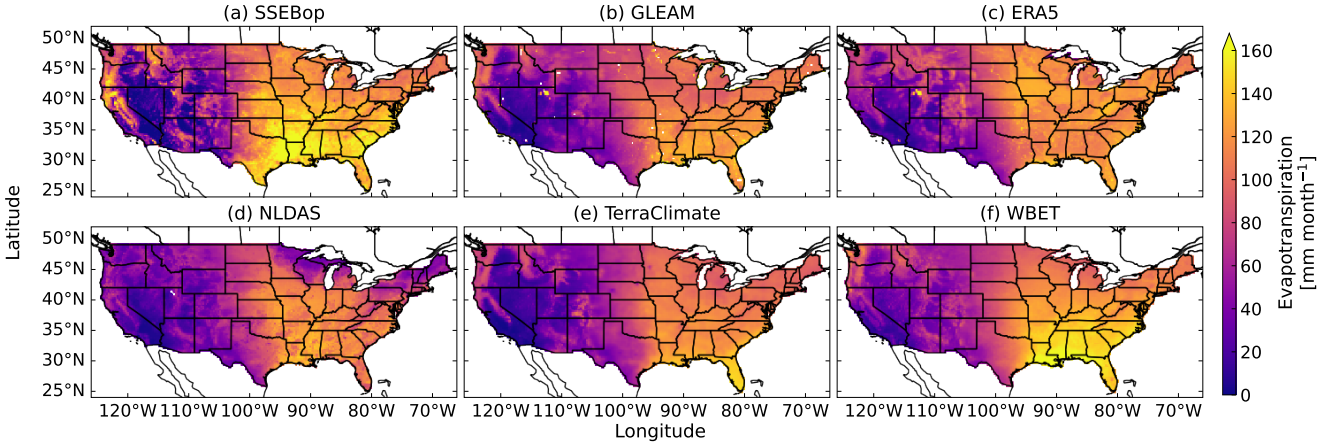

Figure 2. Mean summer (June, July, and August) evapotranspiration (ET) estimates for the Conterminous United States (CONUS)-wide (a) SSEBop, (b) GLEAM. (c) ERA5, (d) NLDAS, (e) TerraClimate, and (f) WBET data sets collocated on the GLEAM grid. These maps clearly show the variation in summer time ET estimates between data sets, while also showing some overall common trends of lower ET in western states and higher ET in the gulf and southern coastal states. Base from Natural Earth, 2024.

missing data. We chose to regrid to the GLEAM resolution as trying to increase resolution in a data set would add additional uncertainty to that data set, because increasing resolution attempts to add information that does not actually exist. Instead, decreasing resolution via a conservative method prevents any increase in uncertainty. As an example of the regridding result, Figure 2 shows the mean summer ET (summer being the months of June, July, and August) for each of the six ET data sets in CONUS collocated on the GLEAM grid.

It is important to note that while regridding makes the data sets collocated, they still might contain "representativeness errors" in the TC method. This type of error may exist if the spatial scale or physical state (representativeness) is highly different between the data sets, which can cause small cross correlations between the errors (i.e., $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) > 0$; McColl et al., 2014; Gruber et al., 2016b). In our case in spatial terms, the ET data sets all originally had their own native resolution, which we decreased to match the GLEAM resolution. So, it is possible that data sets with high native resolution may be penalized for not resolving coarse-scale features in the data sets with lower resolution (see Appendix B of Gruber et al., 2016b). Additionally, the physical representativeness (i.e., the ET state measured) is likely not exactly the same between data sets. Since each data set uses different methods to estimate ET, it is possible that the different methods may include certain ET components that others do not (e.g., condensation in ERA5). Therefore each data set may not be calculating the "same" ET. We discuss the effect of representativeness errors on the overall error variances in Section 5.3.

3.8 Regional Aggregation

Since having accurate and consistent ET data is critical to hydrologic modeling for water availability assessments, we regionally aggregated each ET data set over three regions of concern to water resource management: the Central Valley, High Plains

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

aquifer region, and Upper Colorado River Basin (UCRB), in order to evaluate the agreement between the ET data sets in these regions. We selected the Central Valley and High Plains aquifer regions as they are regions of critical concern to water resource management. Together the two regions account for almost half of irrigated land use and approximately 40% of all groundwater use in the United States (Dennehy et al., 2002; Faunt, 2009). Additionally, both regions have been experiencing drastic declines in aquifer storage, exacerbated by recent droughts. As for the UCRB, the melting winter snowpack in the spring is its main source of water. This influx of water in the spring allows for the region to account for another 5% of irrigated land use in the United States, with 75% of the beneficial water usage going to agriculture as estimated by the Bureau of Reclamation. However, recent low snowpack accumulation and increased water demand has led to a shortage of water in two major downstream reservoirs, Lake Powell and Lake Mead, which supply the majority of the water to the lower half of the river basin. Therefore, as these three regions account for over 50% of the irrigated land in the United States, evaluating the accuracy and consistency between utilized ET data sets (which is a measure of agricultural consumptive water use by crops) can help water resource managers better select an appropriate ET data set when assessing agricultural water use.

For our study, we used the boundaries for each region as derived by the U.S. Geological Survey (U.S. Geological Survey, 2023b, a, 2024). As an initial inspection for trends in ET for these regions, we spatially aggregated each ET data set using an area-weighted average. These regional aggregations resulted in monthly time series data spanning the full date range of each data set. These time series, limited to the common date range between data sets along with the geographic extent of each region, are shown in Figure 3. From a qualitative perspective, it can be seen that data sets appear to be in good agreement in the High Plains and UCRB regions. However, in the Central Valley, there appear to be large inconsistencies between data sets, with some estimating ET that is almost double to triple that of others at the yearly peaks. Additionally, the peak ET timing varies between data sets, with SSEBop and WBET having their peaks shifted later in the year by two to three months compared to the other four data sets. A quantitative agreement comparison and explanation for the inconsistencies in the Central Valley are presented in Sections 4.2.2 and 5.2, respectively.

4 Results

4.1 Collocation Error Variances and Covariances

4.1.1 CONUS Estimates

With the data sets collocated, we applied EC on a pixel-by-pixel basis to estimate the error covariance matrices between each of the data sets. The EC error estimates were calculated for every possible combination of data set quadruplets assuming two were cross-correlated and all the EC assumptions were met. In Section 5.3, we assess and discuss if our data sets meet the EC assumptions and the implications on the error estimates if they are not met. As the agreement probability only uses the pairwise error variances and covariances, we extracted the cross-correlated portion of the error covariance matrix and discarded the uncorrelated error variance estimates[7]. Specifically, by selecting two data sets to be cross-correlated, there are

[7]The uncorrelated error variance estimates are redundant, as they are calculated multiple times in a quadruplet with a cross-correlated pair.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

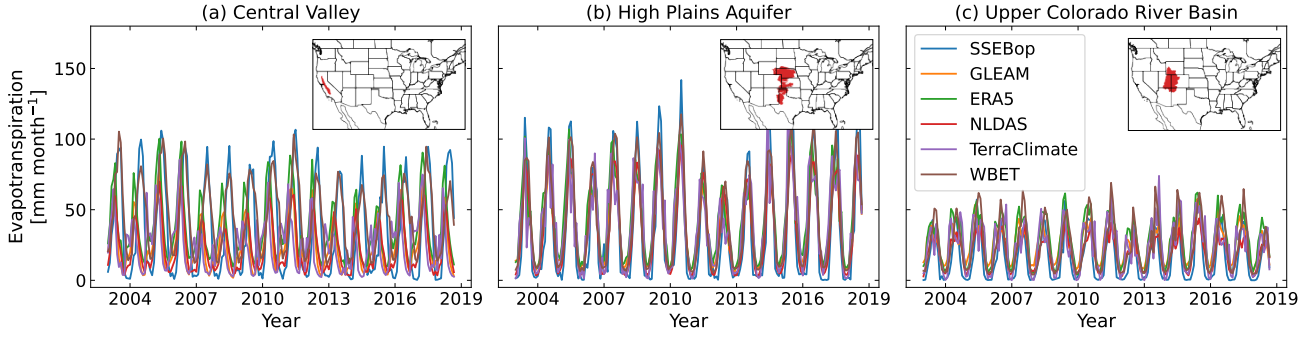

Figure 3. Monthly time series of evapotranspiration (ET) for each data set spatially aggregated over (a) the Central Valley, (b) High Plains aquifer, and (c) Upper Colorado River Basin using an area weighted average within each region. The time series span the common date range between the ET data sets of January 2003 to September 2018. The location of each region in the Conterminous United States (CONUS) is shown in red within an inset in the upper right of each time series panel.

fifteen combinations ($\binom{6}{2} = 15$) of data set pairs. Then, selecting two more to be uncorrelated out of the remaining four to make a quadruplet for EC gives six ($\binom{4}{2} = 6$) estimates for the extracted cross-correlated portion of the error covariance matrix. Therefore, we have six estimates of the extracted cross-correlated portion of the error covariance matrix for each of the 15 cross-correlated pairs.

Although we limited our EC application to quadruplets, EC could easily be applied to sets of five or all six of our data sets to estimate multiple error variances and covariances simultaneously. However, this would not change the calculated covariance matrix estimates compared to using quadruplets as the error variances (covariances) estimated from Equation 4 (Equation 7) only use three (four) data sets at a time. Therefore, we limited the computation to quadruplets to simplify the conceptualization of which data sets are considered cross-correlated and which are considered independent.

When applying EC to each combination, we limited the data sets to the common date range of the quadruplet to maximize data usage. Potentially, this variation in date ranges could cause the data sets with larger date ranges (e.g., ERA5 and WBET) to have smaller variation in their six error covariance matrix estimates than the data sets with more recent data ranges (e.g., SSEBop and GLEAM). However, we found that this effect of variation in date ranges did not occur for these or any of the data sets. In fact, when comparing the fractional differences in error variances from the extracted covariance matrices for the possible combinations of a given data set with the mean of those combinations, we found that at least 60% (86%) of the values for all combinations are within a fractional difference of 0.25 (0.5) of the mean error variance. Additionally, the quantile distributions of the fractional differences are highly symmetric for each data set with the median close to a zero fractional difference. Therefore, we concluded that the variation in date ranges is not having a significant effect on the error variance estimates, and therefore not using a common date range across all data sets is acceptable.

To show the extracted covariance matrix estimates in a simplified way, we computed the error cross-correlations with Equation 7 for each data set combination. In Figure 4, the median of the six error cross-correlations estimates for the fif-

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

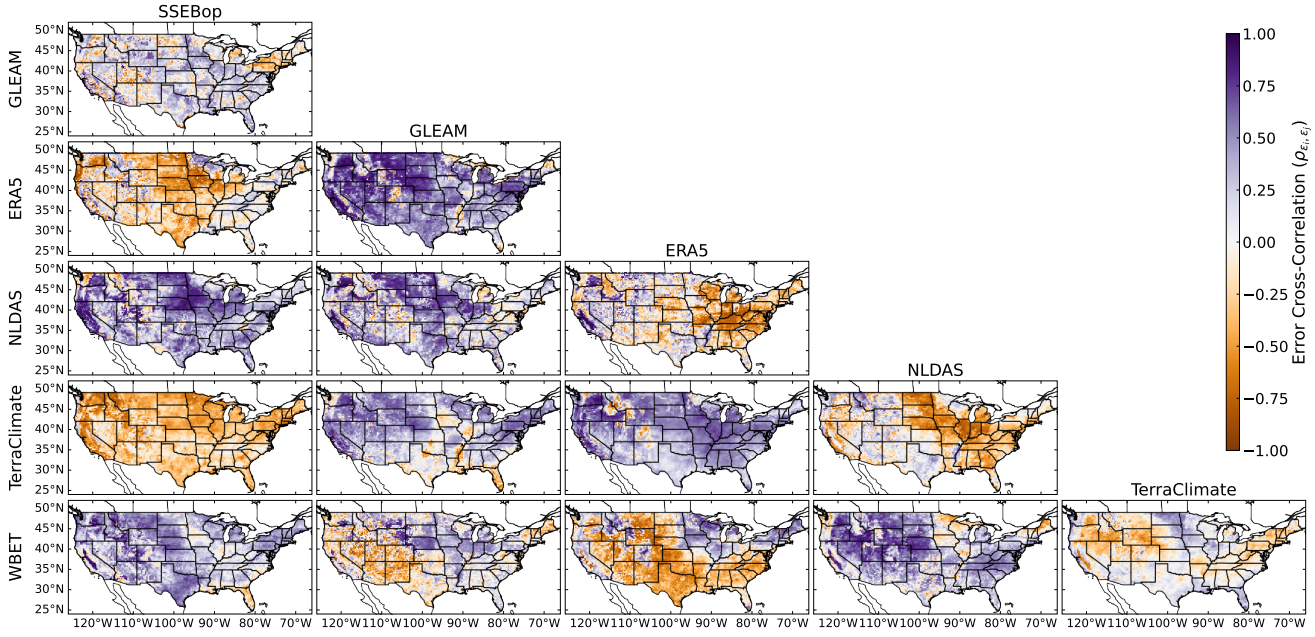

Figure 4. Median extended collocation (EC) error cross-correlations of the six EC error estimates for each evapotranspiration (ET) data set pair. The figure panels are labeled with one data set in the pair being given above the columns and the other data set given next to the rows. Characteristics of the six ET data sets are provided in Table 1. Base from Natural Earth, 2024.

teen correlated data set pairs are shown. These maps show a large variation in the spatial patterns of the error cross correlations across CONUS, with some data set pairs having practically all positive (e.g., GLEAM-ERA5) or negative (e.g., SSEBop-TerraClimate) correlations across CONUS while others have a mix or generally low values (e.g., SSEBop-GLEAM or TerraClimate-WBET). Unlike the error variances, which had low variation between the estimates for a given data set, the error cross-correlations vary greatly between the six estimates, with some indicating strong positive correlations and others indicating zero to negative values, making it impossible to know whether the datasets are or are not cross correlated. However, several data set pairs, like ERA5-TerraClimate, have highly consistent estimates with positive correlations, indicating that these data set pairs have errors that are cross-correlated. Therefore, confirming the presence (or lack) of error cross correlations between data sets from the EC method is possible if multiple data set quadruplets show consistent correlation results.

Besides applying EC to the full record within the common date ranges, we also separated the months in each year into their respective season (e.g., winter included December, January, and February) and applied EC to the seasonal data sets, as ET has obvious seasonal trends. We note, however, that separating the data by seasons also reduces the amount of data used for the EC error covariance matrix estimation, which will result in less accurate error estimates (Zwieback et al., 2012). Like the full record cross correlations presented above, seasonal cross correlations values presented large variations between the six covariance matrix estimates as well as spatial variations between seasons. This is expected as the decreased amount of monthly

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

data used for each season will increase the variation in the error covariance matrix derived from the EC method (Zwieback et al., 2012) in addition to ET having different spatial patterns in each season.

4.1.2 Regional Estimates

To estimate the error covariance matrices for the three regions of concern, we spatially aggregated the gridded variance and covariance components of the matrices from Section 4.1.1 using an area-weighted average across each region's bounding area. Alternatively, we could have recomputed them using the regional time series derived in Section 3.8, as these time series can be considered collocated. However, we chose the former method for two reasons. The first reason is that the difference between the two methods is the order in which to average the spatial and temporal dimensions. As we wanted to retain the spatial variation over temporal variation, we chose the former method as it performs the spatial average last. The second reason has to do with the collocation analysis method itself. As with any variance (or covariance) estimate from randomly sampled data, the variation in the variance estimate will decrease with increasing data. In terms of collocation analyses, if the amount of data is small enough that the variation in an error covariance matrix is larger than the true variance itself or the variances of the other data sets' errors, it is possible to obtain covariance matrices that are not semi-positive definite. This is typically more of a problem for data sets with low error variances, especially when grouped with data sets that have higher error variances. Relating this back to our spatial and temporal averaging order, numerous grid cells in the CONUS EC covariance matrices had non-semi-positive definite structure. This is also the case for some estimates if we recomputed the values from the regional time series. Since these non-semi-positive definite matrices are nonsensical, we discarded them and treated them as missing data. Therefore, by choosing to spatially aggregate the CONUS EC results, we can spatially average over the regions ignoring missing data, which is not possible if we used the regional time series data.

In Figure 5, the results of this spatial averaging are shown as the error cross correlations for the three regions. Since there are multiple estimates of the correlations due to the six possible data set combinations, each with differing amounts of missing data, the values shown are the weighted mean of these estimates with the weights being the fraction of regional area associated with non-missing data, normalized across the different estimates. This weighted mean allows for a clearer representation of the values, by giving each estimate equal weight based on its area of non-missing data. To give an idea of the variation from the different estimates, error bars are included, which show the unweighted 16th and 84th percentile range of the estimates. From the figure, it can be seen that the High Plains Aquifer and UCRB have relatively low correlations ($\rho < 0.5$), while the Central Valley has approximately half of the data set pairs with high correlations ($\rho > 0.6$). Across all regions, ERA5-TerraClimate and SSEBop-WBET show consistent positive correlations, whereas SSEBop-TerraClimate show negative correlation, which can also be seen in Figure 4. This consistency between the CONUS estimates and regional aggregation strongly indicates that these data set pairs do in fact have correlated errors.

As we did with the CONUS estimates, we investigated the seasonal error cross correlations by applying the same regional aggregation process to the seasonal CONUS estimates. Trends in the seasonal cross correlations show that the values still presented large variations between the six estimates as well as between the seasons. This variation is apparent with >40% of the data set pairs across regions and seasons having 16th and 84th percentile ranges that include $\rho = 0$. Again, this variation

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

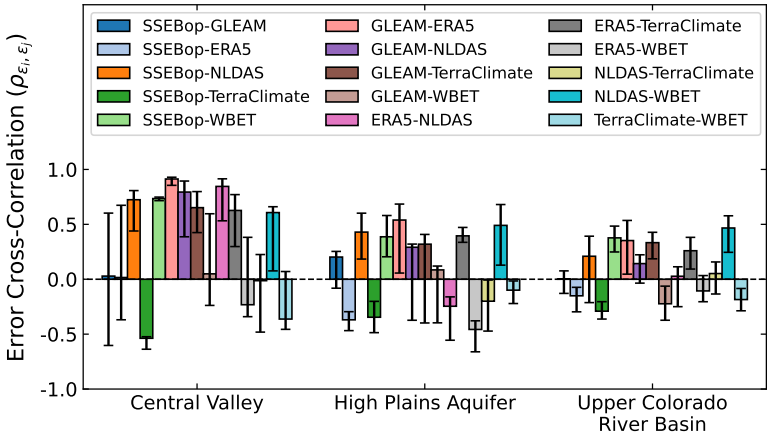

Figure 5. Regionally aggregated extended collocation (EC) error cross correlations for each evapotranspiration (ET) data set pair. The plotted values are the weighted mean of the different data set pair estimates with the weights being the fraction of regional area associated with non-missing data, normalized across the different estimates. The error bars show the unweighted 16th and 84th percentile range of the different combination estimates. Characteristics of the six ET data sets are provided in Table 1.

is expected due to the decreased amount of monthly data by restricting to a season and the different ET spatial patterns across each region.

4.2 Statistical Agreement Evaluation

4.2.1 CONUS Estimates

To estimate the agreement probabilities across CONUS, we first calculated the absolute relative bias between each data set pair for every element in their time series at each grid cell. We then calculated the temporal median of the absolute relative bias for each grid cell limited to the common date range used in the EC calculations for consistency with the EC error estimates. We note that by limiting the temporal median of the bias to the EC date ranges, this resulted in six estimates of the relative bias per data set pair, since four data sets were used in the EC calculation ($\binom{4}{2} = 6$). As the only difference in these six median biases for a given data set pair are the date ranges, the biases are highly consistent, with the median differences between the six estimates for all data set pairs across CONUS being <3%. However, we still chose to use different data ranges within a data set pair for consistency with the EC error estimates.

Using the EC error covariance matrix estimates from Section 4.1.1 and these biases, we calculated the agreement probabilities using the methods described in Section 2.2. In Figure 6, the median agreement probabilities of the six estimates per data set

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Table 2. The percentage of grid cells across the Conterminous United States (CONUS) with agreement probabilities above the given significance levels for each of the evapotranspiration (ET) data set pairs. Characteristics of the six ET data sets are provided in Table 1.

Data Set Pairs	Non-seasonal			Winter			Spring			Summer			Fall		
	Significance Levels			Significance Levels			Significance Levels			Significance Levels			Significance Levels		
	0.16	0.05	0.01	0.16	0.05	0.01	0.16	0.05	0.01	0.16	0.05	0.01	0.16	0.05	0.01
SSEBop-GLEAM	71.2	87.7	93.8	8.2	28.3	50.2	32.9	54.4	71.6	32.8	58.2	79.3	44.8	64.9	78.9
SSEBop-ERA5	79.2	93.9	97.6	15.5	38.1	62.6	34.1	55.9	75.1	49.9	71.0	84.7	43.5	68.5	82.8
SSEBop-NLDAS	79.3	91.6	96.3	42.6	60.8	77.2	50.3	72.4	84.6	30.7	48.2	63.0	57.7	75.4	86.6
SSEBop-TerraClimate	94.1	98.4	98.6	61.5	77.5	85.9	39.0	69.7	90.8	61.5	82.1	92.0	56.1	87.4	97.2
SSEBop-WBET	87.3	97.1	97.8	37.2	54.0	68.0	55.0	81.5	92.5	62.8	81.8	92.2	60.5	77.1	90.3
GLEAM-ERA5	84.0	95.4	97.3	63.2	80.2	89.7	35.0	58.2	74.8	56.6	72.4	85.1	67.1	87.1	94.9
GLEAM-NLDAS	65.0	79.8	89.3	17.7	39.6	63.5	22.6	38.0	55.3	45.1	62.9	75.8	55.3	71.8	80.5
GLEAM-TerraClimate	92.0	97.4	97.6	37.7	61.1	75.1	58.9	77.5	87.7	80.7	91.0	94.2	81.5	95.5	97.2
GLEAM-WBET	85.8	94.6	96.1	27.9	49.6	68.0	29.6	54.2	73.6	44.7	76.0	89.8	75.9	86.9	93.5
ERA5-NLDAS	70.9	89.8	94.6	29.6	50.5	69.5	22.1	34.4	50.1	54.4	70.0	79.7	40.8	69.3	85.6
ERA5-TerraClimate	96.0	98.4	98.6	41.7	64.1	80.5	70.2	90.1	96.3	67.3	87.0	93.4	88.8	97.0	98.1
ERA5-WBET	91.6	96.6	97.4	33.6	56.9	73.6	26.3	47.3	75.3	59.2	84.8	94.1	69.2	86.1	93.0
NLDAS-TerraClimate	72.3	93.6	97.3	43.3	57.4	70.5	35.7	53.4	64.4	65.1	78.8	87.3	58.5	80.9	91.5
NLDAS-WBET	93.6	96.9	97.2	58.2	80.7	91.8	76.3	92.8	96.6	43.0	64.9	76.3	83.4	94.9	96.8
TerraClimate-WBET	94.5	97.8	97.8	34.3	51.1	61.0	40.1	69.1	87.3	59.2	86.3	95.7	66.7	90.4	97.3

pair are shown for all of CONUS[8]. These plots show that >70% of grid cells for all data set pairs agree across CONUS (except for GLEAM-NLDAS) at a conservative significance level of 0.16, with disagreement primarily occurring in the mountainous West and the Southeast (see Section 5.1). Lowering the significance level to 0.05, >85% grid cells for all data set pairs (again, except for GLEAM-NLDAS) have agreement probabilities above this level. A table detailing the percentage of grid cells above significance levels of 0.16, 0.05, and 0.01 for each of the data set pairs are given in Table 2.

As ET data have strong seasonal trends, we also assessed the agreement probabilities across seasons by limiting the bias estimates to the matching months as used in the seasonal error covariance matrix estimates in Section 4.1.1. Figures similar to Figure 6 for the seasonal agreement probabilities can be found in Appendix A. These figures show that winter has the lowest agreement probability across CONUS, with a median of 0.073 across all grid cells and data set pairs. Winter is followed by spring (median across grid cells and data set pairs of 0.095), then summer (0.180), and fall (0.232). In terms of agreement from a 0.16 significance level (with agreement at a 0.05 significance level written in parentheses), winter has 36.8% (56.7%) of grid cells across data set pairs above the significance level, with spring, summer, and fall having 41.9% (63.3%), 54.2% (74.4%), and 63.3% (82.2%), respectively. Percentages for each data set pair are also given in Table 2. Therefore, these seasonal results

[8]We note that we used the median primarily to account for and ignore missing data in the error covariance matrices (see Section 4.1.2). However, this choice will have minimal influence on any interpreted results as the six estimates, like the biases, are highly consistent, with 75% of grid cells across estimates and data set pairs differing from their median probability by <0.01 (and 99% of grid cells differing from their median probability by <0.06).

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

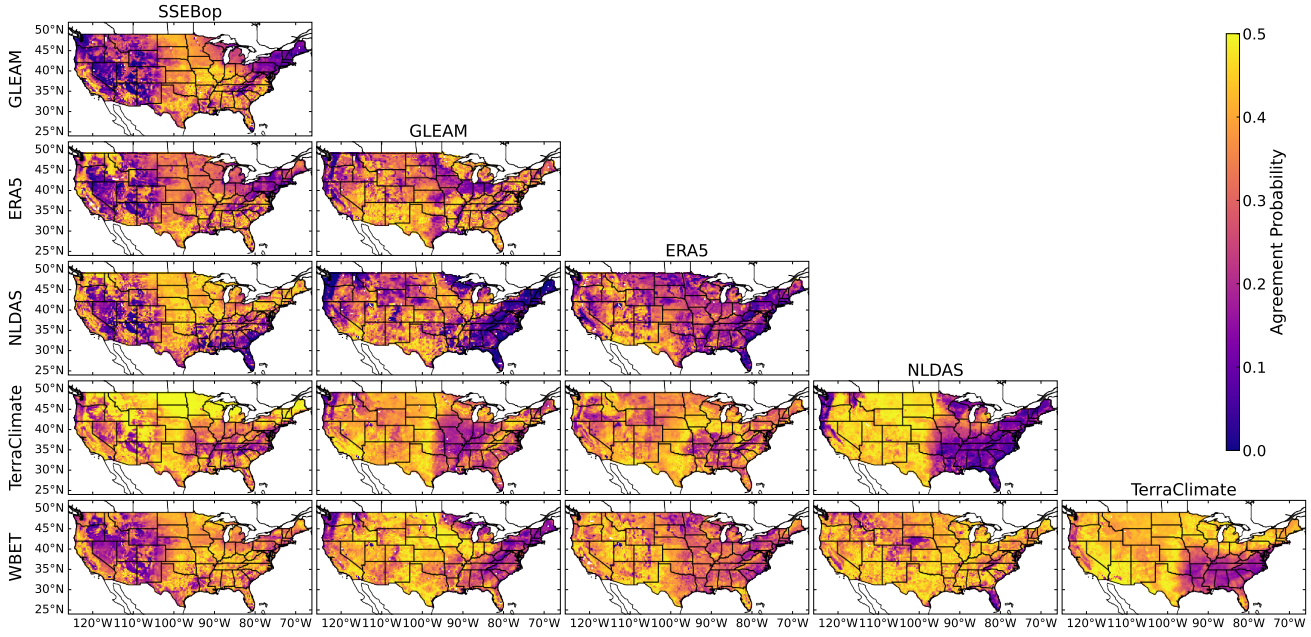

Figure 6. Median agreement probability grids of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair. The figure panels are labeled with one data set in the pair being given above the columns and the other data set given next to the rows. These maps show that overall most data set pairs are in excellent agreement, with some low agreement probabilities occurring between various pairs in the mountainous West or Southeast. Characteristics of the six ET data sets are provided in Table 1. Base from Natural Earth, 2024.

show that, while summer and fall have reasonable agreement between pairs, the high probabilities from the full record results do not carry over to the seasons. The reason for this is discussed in Section 5.1.

4.2.2 Regional Estimates

To estimate the agreement probabilities for the three regions of concern, we spatially aggregated the gridded temporal medians of the absolute relative biases from Section 4.2.1 using an area-weighted average across each region's bounding area. As discussed in Section 4.1.2, we could have first aggregated the biases spatially and then temporally, but we chose temporal then spatial aggregation to prioritize the spatial variation rather than the temporal variation. Then, using the regional error covariance matrix estimates from Section 4.1.2, we computed the agreement probabilities for each region of concern. The median agreement probabilities over the six-combination estimates are shown in Figure 7 along with error bars indicating the 16th to 84th percentile range of the estimates. From this figure, it can be seen that the six agreement probability estimates for the data set pairs are very consistent (i.e., small percentile ranges). This is a promising result as it shows that the variation in covariance matrices is not having a major influence on the probabilities, thereby indicating that any possible violations of the

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

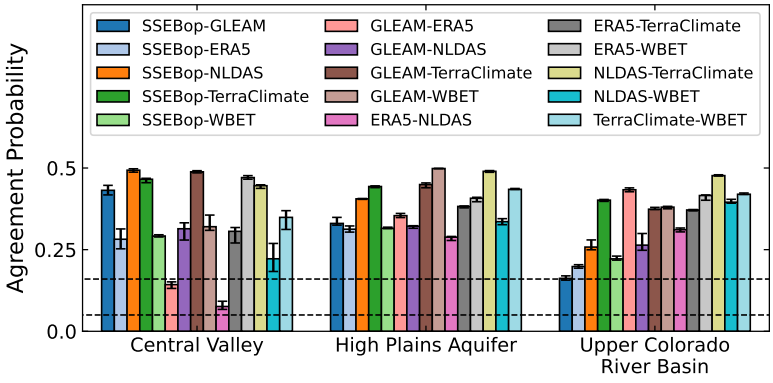

Figure 7. Median regional agreement probabilities of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair, with error bars indicating the 16th and 84th percentile range of the different estimates. The bar chart shows that all but two data set pairs are in agreement based on a 0.16 significance level (upper dashed black line). The two pairs that are below this significance level, but above a 0.05 significance level (lower dashed black line), are ERA5 paired with GLEAM or NLDAS in the Central Valley. Characteristics of the six ET data sets are provided in Table 1.

EC assumptions in our application do not significantly impact our results. In terms of median values, all regions have high probabilities indicating agreement between all data sets at a conservative significance level of 0.16, except for the two data set pairs of GLEAM-ERA5 along with ERA5-NLDAS in the Central Valley. However, these two data set pairs do show agreement when comparing to the less conservative 0.05 significance level. Therefore, just like the gridded probabilities for all of CONUS, the data sets show no statistical disagreement when looking at the full record.

Finally, to assess the agreement probabilities across seasons, we spatially aggregated the seasonal median biases and EC covariance matrices to compute seasonal probabilities for each region, which are shown in Figure 8. These seasonal results show, similarly to the CONUS seasonal probabilities, that there is excellent agreement (i.e., consistently high probabilities) as we observed for the CONUS full record results. Starting with the region with the highest overall seasonal agreement, the High Plains, the summer and fall have all data set pairs with p-values > 0.1, indicating reasonable agreement. As for the winter and spring, only two data set pairs fall below the 0.05 significance level in each season, which is good considering the median gridded p-values across CONUS were 0.073 and 0.095 for winter and spring, respectively. Looking at the UCRB, only the summer has all data set pairs with p-values above 0.05, with the other seasons having 2–4 pairs of data sets with values below 0.05. However, it is important to note that most of these pairs are combinations with SSEBop, which has been shown to not perform as optimally (i.e., it overestimates ET) in mountainous regions with elevations greater than $2,500$ m like the UCRB (Senay, 2018). Finally, the Central Valley shows similar trends in probabilities in the spring and fall as the UCRB, with 1–3 data set pairs having p-values below 0.05. Yet, when looking at the winter and summer, only a few data sets have p-values > 0.16, with the majority of data set pairs having extremely low agreement (p-values < 0.015). This is especially true in the summer, where 8 pairs (SSEBop or WBET paired with GLEAM, ERA5, NLDAS, or TerraClimate) have p-values < 0.005.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

This stark disagreement between multiple data sets in the summer stems from the large inconsistencies (i.e., peak ET value and month of occurrence) that were found in the regional time series shown in Figure 3. As these disagreements in ET data sets could have implications for hydrological modeling in the Central Valley, we discuss the reasons for the inconsistencies in detail in Section 5.2.

5 Discussion

5.1 CONUS Agreement

Our results of the gridded agreement probabilities across CONUS in Section 4.2.1 showed that most data set pairs have high agreement probabilities. Areas with consistently low probabilities across multiple pairs were found to be linked to SSEBop in the mountainous West or TerraClimate in the Southeast. As stated above, the low agreement probabilities for SSEBop in high, complex mountainous regions are expected, as SSEBop is known to overestimate ET at elevations > 2500 m (Senay, 2018). While this is not seen in Figure 3 for the UCRB, which has higher elevations, the overestimation is likely suppressed by the spatial aggregation, which averages out the extremes occurring in SSEBop in this region as seen in Figure 2. As for TerraClimate, Liu et al. (2023) found that in a comparison of ten global ET products at site and basin scales, SSEBop and TerraClimate both showed relatively poor results compared to the other data sets. Specifically, they also found that TerraClimate had inconsistent ET estimates in the "temperate" regions like the southeast of CONUS. The reasoning for this poor agreement in our and their studies likely originates from the input data products or water balance model used to generate the ET in the TerraClimate data set. For example, a 1-D soil water balance model like TerraClimate is essentially assessing the climatic water deficit (the difference between precipitation and atmospheric demand) to derive ET. By not including any component to account for the state of vegetation on the surface, TerraClimate seems to perform less optimally in "temperate" regions that are highly vegetated like the southeast of CONUS. As for the GLEAM-NLDAS pairs having low agreement across CONUS, it is due to the two data sets having the smallest error variance estimates out of all six data sets. Therefore, it is not unexpected that this pair would have lower agreement probabilities as smaller errors would narrow the bias distribution compared to the larger errors of, for example, TerraClimate.

Beyond the full record results, the same general patterns of low agreement probability areas are also present in the seasonal results. However, one thing of interest from the seasonal results is how the CONUS probabilities are generally less than those of the full record results. The reason for this increase in the probabilities for the full record data is due to the strong seasonal trends in ET. Specifically, when computing the full record agreement probability, the less biased winter data (i.e., ET is typically lowest in the winter across CONUS due to being energy limited) decreases the overall temporal bias, while the more uncertain summer data (i.e., larger error variances) increases the bias error. Hence, the bias distribution in the agreement probability calculation is more closely centered on 0 and wider, resulting in larger probabilities. In contrast, the seasonal results do not have both effects of bias dampening and increased errors, rather they have one effect or the other. For example, even though winter is less biased, it has very small errors due to less variability. Alternatively, summer was found to have larger errors, but it

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

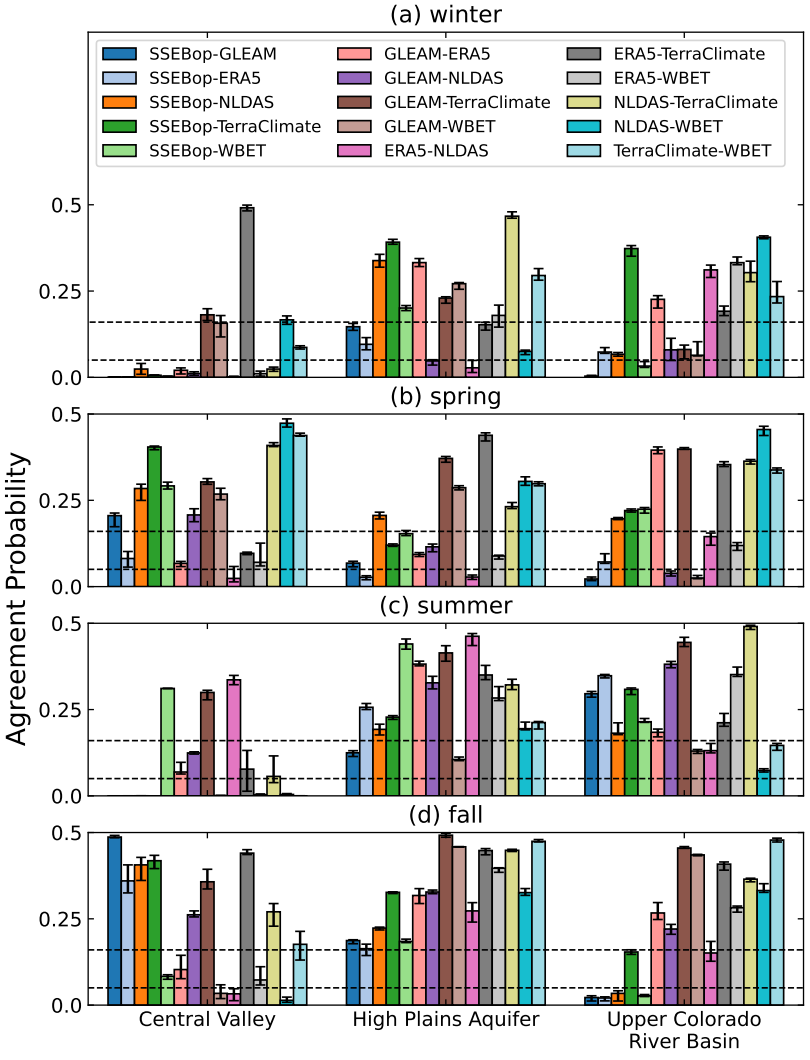

Figure 8. Median regional agreement probabilities of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair separated by (a) winter, (b) spring, (c) summer, and (d) fall. Error bars indicate the 16th and 84th percentile range of the six different estimates. The bar charts show that data set pairs in the High Plains and UCRB are typically in agreement for all seasons based on a 0.05 significance level (lower dashed black line). However, the Central Valley shows that, while most data set pairs agree at the 0.05 significance level in the spring and fall, the majority of data set pairs have probabilities well below 0.05 in the winter and summer. Characteristics of the six ET data sets are provided in Table 1.

510 also can have large biases due to the increase in ET. Therefore, we found it imperative when applying our statistical agreement method to highly seasonal data that seasonal results should also be generated in addition to any non-seasonal results.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

5.2 Regional Agreement

Like the CONUS agreement probabilities, the regional agreement probabilities in Section 4.2.2 showed that most data set pairs have high agreement probabilities (p-values > 0.16) for the full record data. When separated by season, the results showed that the High Plains and UCRB maintained relatively high probabilities across all seasons, with the UCRB showing some low probabilities with SSEBop pairs due to the region's high average elevation. However, while the Central Valley showed agreement probability levels similar to the High Plains and UCRB in the spring and fall, probabilities were found to be extremely low in winter and in the peak ET season of summer. Specifically for the summer, SSEBop and WBET showed strong disagreement (i.e., p-values < 0.005) when paired with the other four data sets (GLEAM, ERA5, NLDAS, and TerraClimate). This inconsistency between the data sets in the summer originates from the large variations in peak ET values along with the temporal shift for when this peak occurs as discussed in Section 3.8 and shown in Figure 3.

Specifically, the peak ET month for GLEAM, ERA5, NLDAS, and TerraClimate in the Central Valley is mid-spring (approximately April for most years), whereas the peak ET for SSEBop and WBET is mid-summer (approximately July). Typically, ET is expected to peak in mid-summer if the observed area is not water limited, as this is when temperatures and net radiation and thus potential evapotranspiration are the highest. As this is not the case for GLEAM, ERA5, NLDAS, and TerraClimate, it implies that these ET data sets are water limited in the Central Valley. Looking at general precipitation trends for the region, precipitation primarily occurs in the winter months and is minimal in the summer. So, based on precipitation patterns alone, it would be expected that the Central Valley is water limited in the summer, and ET should peak in the spring. However, the region has major agricultural production, and to sustain this production in the summer, the region is highly irrigated. Thus, the water input from irrigation means that ET should actually peak in the summer.

In terms of data sets, ERA5, NLDAS, and TerraClimate explicitly state that they do not include an irrigation component, while WBET includes groundwater-sourced irrigation in its water balance model. As for SSEBop and GLEAM, irrigation is not included as an input like it is in WBET, but the models attempt to account for it. For SSEBop, which was designed for operational use in irrigated regions, irrigation is implicitly accounted for in the ET calculations via the temperature difference between hot and cold boundary conditions that arises from evaporative cooling. GLEAM also accounts for irrigation implicitly by using satellite soil moisture data as an input. However, it appears that the soil moisture data used in GLEAM is not properly accounting for the irrigation, likely due to the relatively arid environment of the Central Valley which dampens any increase in soil moisture signal from irrigation at GLEAM's relatively low spatial scale. Therefore, we conclude that the inconsistencies found between the six data sets in the Central Valley can be directly attributed to irrigation. This includes both the temporal shift and increase in peak values for SSEBop and WBET, which not being water limited, can have peak ET approximately 1.5–2 times higher than the other four data sets. As a result of this disagreement, any hydrological modeling of the Central Valley could produce biased results based on the choice of ET data set if the data set does not properly account for irrigation.

Finally, this leads to the question of why the High Plains does not show these inconsistencies between data sets, since it is also a highly irrigated region. The reason for this is that the monthly precipitation patterns in the High Plains are not the same as the Central Valley, as the High Plains typically continues to receive rain throughout the summer. While these rains

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

are intermittent and thereby crops require supplemental irrigation, monthly precipitation does occur on average. Therefore, a continual supply of water from precipitation throughout each month prevents the ET estimates from being water limited as they were in the Central Valley without irrigation.

5.3 TC Assumptions

One caveat of this study is the use of the collocation analysis methodology and assuming the utilized data sets met all of the required assumptions of the method. Starting with the first assumption of stationarity, it is possible that our signal and random errors are not stationary. In terms of our ET data, it is obviously not stationary with the clear presence of seasonality, and it is a reasonable assumption that the seasonal pattern is similar between data sets at most collocated grid cells. While non-stationarity may seem like an issue, in reality it is not if the data sets all have the same non-stationarity effect. Therefore, stationarity in the signal is likely only a minor, if not insignificant, issue for the accuracy of the EC error estimates (and subsequent agreement probabilities), as is commonly found for most applications (Gruber et al., 2016b).

As for the stationarity of the errors, it requires that the error estimates remain constant throughout the common date range. However, as we found in Section 4.1, the derived errors are different between seasons, with winter having much lower values than the summer. This indicates, as we concluded in Section 5.1, that seasonal results should be generated in addition to any full record results. Without the seasonal separation, the full record, non-seasonal error variances are averaged over the entire data range, which can limit their representativeness when trying to apply them during a specific season.

For the next assumption of error cross-correlation, it was clear from the results in Section 4.1 as shown in Figure 4 that some data set errors are indeed cross correlated for the majority of CONUS (e.g., GLEAM-ERA5, ERA5-TerraClimate). Interestingly, these cross-correlated data sets each utilize a different reference data type and/or calculation method. As stated in Section 3, it is commonly assumed that the errors between data sets with different reference data type and/or calculation methods should have minimal cross correlations, while those with matching reference data type and calculation method are more likely to have correlated errors. Hence, we would have expected SSEBop-GLEAM to be correlated along with ERA5-NLDAS, and TerraClimate-WBET, rather than e.g. GLEAM-ERA5 and ERA5-TerraClimate as we found. Therefore, this TC application demonstrates that assuming different reference data type and/or calculation methods equates to independent errors is not necessarily correct. Verifying this finding could be done by estimating error cross-correlations between data sets and flux tower data similar to the approach in Volk et al. (2023) to find how much error depends on the true ET value from the flux tower (despite the scalar mismatch). However, performing this verification is beyond the scope of this paper.

One thing to note about the potential inclusion of positively cross-correlated data sets is that the resulting error variances will be underestimated (Yilmaz and Crow, 2014). Therefore, the EC error estimates presented in this paper should be taken as the lower limit of the random errors. In terms of agreement probabilities, a random error lower limit means that the corresponding probability is also a lower limit. So, the agreement probabilities presented in Section 4.2 could actually have higher values, implying any disagreement between data sets is not guaranteed.

Compared to the presence of error cross-correlations, error orthogonality and autocorrelation are expected to have a minimal influence on the resulting EC error estimates. For the orthogonality assumption, the change in magnitude of the estimated errors

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

with season is causing the assumption to not be met. However, its effect is negligible if it is approximately equal for at least two of the three "independent" data sets in EC (Yilmaz and Crow, 2014; Gruber et al., 2016b). As the same seasonal variation in the errors is present in all data sets, we concluded that the error orthogonality assumption, while not met for our data sets, can be expected to have insignificant impact on the error estimates. As for autocorrelation in the errors, the change in errors with season is likely also causing this assumption to not be met. However, it is expected that a violation of this assumption will not affect the accuracy of the error estimates, but it will increase the variation (i.e., precision) of the error estimates (Zwieback et al., 2012). Therefore, the agreement probabilities should be insignificantly influenced by the violation of these two assumptions.

Finally, while not one of the four core EC assumptions, representativeness errors can bias one or two of the error variances in the "independent" data set triplet as discussed in Section 3.7. In our case, representativeness errors would likely be dominated by the differences in spatial scales rather than physical representativeness, as the estimated ET in each data set is generally representative of the same physical process. However, the biases from spatial differences are likely minimal, if present, in our estimated errors due to the conservative regridding of the data sets. By conservative regridding, we are aggregating the higher resolution data sets to a common lower resolution such that variations caused by local phenomena outside of a collocated grid cell in the higher resolution data are now present by aggregating multiple cells during regridding. If representativeness errors from spatial differences are not completely suppressed in the data after regridding, their effects will only result in overestimated error variances (see Appendix B of Gruber et al., 2016b). However, this overestimation should be minimal compared to the underestimation caused by the presence of error cross correlations.

6 Summary

In this work, we developed a novel application of collocation analysis that can be used to evaluate the statistical agreement between collocated data products. Our method utilizes the estimated error covariance matrix from EC and the relative bias between data sets to calculate the probability that the data sets have an absolute relative bias that is within the estimated errors. This probability can then be used as a p-value in a statistical hypothesis test at a chosen significance level to determine if the data sets are not in agreement.

We then applied this method to six gridded ET data sets with CONUS coverage (SSEBop, GLEAM, ERA5, NLDAS, TerraClimate, and WBET), which we regridded to a common resolution. These results showed that >70% of grid cells for all pair combinations of data sets agreed across CONUS (except for the GLEAM and NLDAS pair) at a conservative significance level of 0.16. At a lower significance level of 0.05, >85% grid cells for all pair combinations of data sets had agreement probabilities above this level (again, except for the GLEAM and NLDAS pair). As ET has clear seasonal trends, we also estimated the agreement probabilities of the data sets separated by season. These results showed that winter had the lowest median agreement probability across CONUS, followed by spring, then summer, and fall. In terms of agreement from a 0.16 significance level, winter had 36.8% of grid cells across all data set pairs above the significance level, with spring, summer, and fall having 41.9%, 54.2%, and 63.3%, respectively. In terms of agreement from a 0.05 significance level, winter had 56.7% of

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

grid cells across all data set pairs above the significance level, with spring, summer, and fall having 63.3%, 74.4%, and 82.2%, respectively.

In this analysis, it was found that the locations of regions with low agreement probability were typically consistent between data set pairs. So, we further implemented our method across three regions in CONUS, the Central Valley, High Plains aquifer region, and UCRB, as these regions are of increased concern to water resource management. All three regions were found to have p-values > 0.05 for all data set pairs when looking at the full record data, with all but two pairs in the Central Valley having p-values > 0.16. When separated by seasons, the High Plains and UCRB, typically had all data set pairs with p-values > 0.05 for all seasons. Pairs in the UCRB that had p-values < 0.05 were found to mainly be combinations with SSEBop, which is known to not perform as optimally in mountainous regions like the UCRB (Senay, 2018). As for the Central Valley, the probabilities in the spring and fall were similar to the UCRB, with 1–3 data set pairs having p-values < 0.05. However, in the winter and summer, the majority of data set pairs had extremely low agreement (p-values < 0.015), especially in the summer, where 8 pairs had p-values < 0.005. After investigating the reason for this stark disagreement, we concluded that the inconsistencies found between the data set pairs in the Central Valley could be directly attributed to the lack of an irrigation component in the GLEAM, ERA5, NLDAS, and TerraClimate data sets. Therefore, our novel methodology was able to identify a shortcoming within these data sets, which can be used to guide improvements via an irrigation component in future versions of these data sets.

While this work is currently limited by the lowest resolution data set, future applications of this method could be applied to higher resolution ET data sets like OpenET and its ensemble components (Melton et al., 2022). As OpenET is becoming a key data set in supporting water resource and land management applications in western CONUS, evaluating the agreement within its ensemble components and final product beyond comparing to in situ sites could be a valuable assessment (Volk et al., 2024). However, this may require further temporal and spatial coverage as OpenET only has monthly ET data from 2008 to 2024 (or daily from 2016 to 2024) and CONUS coverage for longitudes $\gtrsim 94°$W. While spatial coverage is not necessarily a problem, the low temporal coverage would increase the variation in the estimated error covariance matrix, thereby making the agreement probability estimates less accurate.

Code availability. All analyses in this paper are reproduced in Jupyter notebooks, which are available at https://doi.org/10.5066/P1VN9ENA (Doore et al., 2024).

Appendix A: Seasonal Statistical Agreement Estimates

Figures A1–A4 show the median agreement probability grids of the six relative bias and extended collocation (EC) error covariance matrix estimates for each data set pair separated by season. As the evapotranspiration (ET) data sets are in monthly time steps, the winter includes December, January, and February; spring includes March, April, and May; Summer includes June, July, and August; fall includes September, October, and November.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

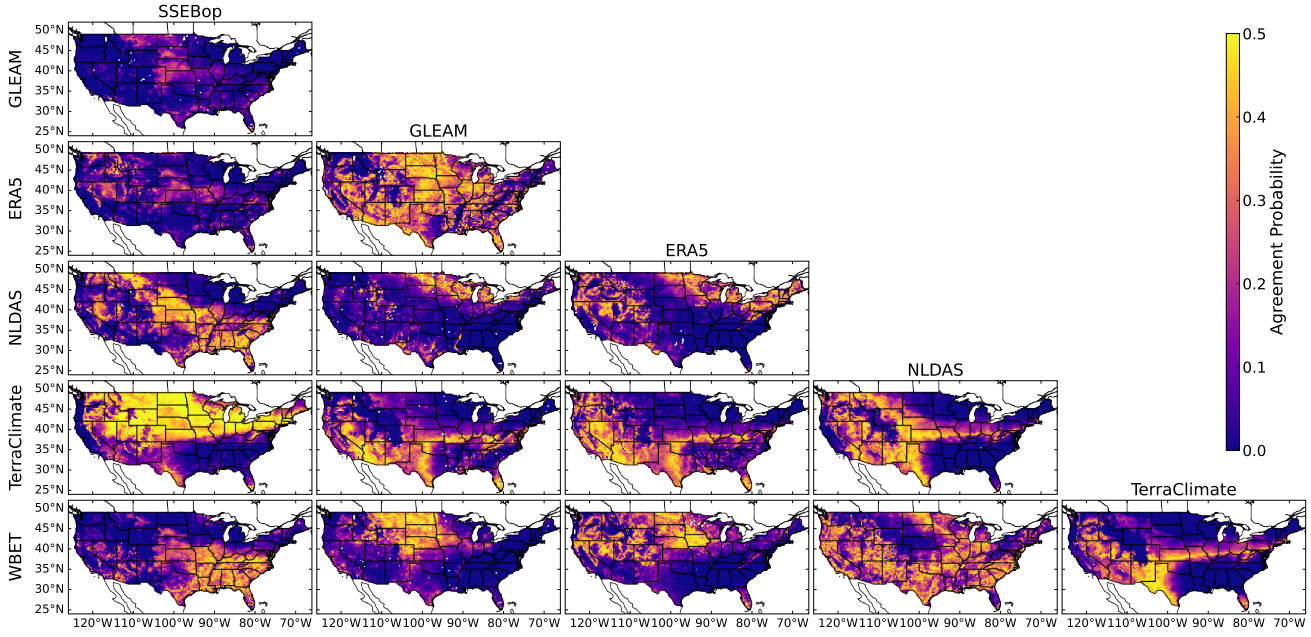

Figure A1. Median agreement probability grids of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair for the winter season. The figure panels are labeled with one data set in the pair being given above the columns and the other data set given next to the rows. Characteristics of the six ET data sets are provided in Table 1. Base from Natural Earth, 2024.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

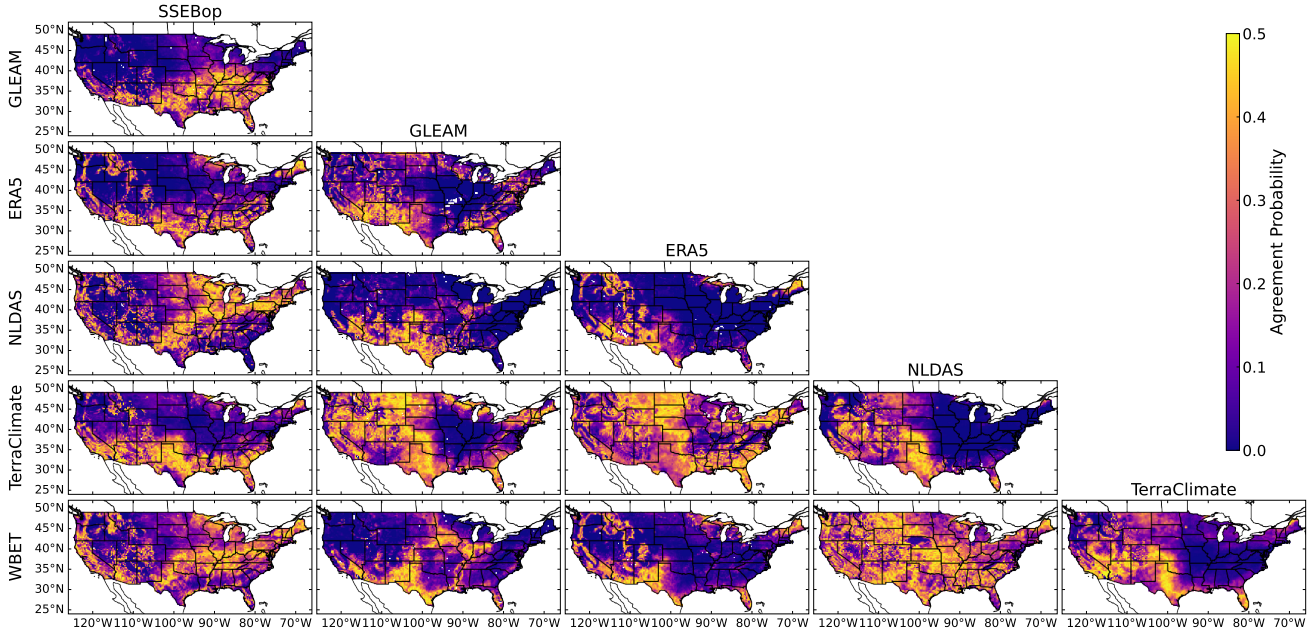

Figure A2. Median agreement probability grids of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair for the spring season. The figure panels are labeled with one data set in the pair being given above the columns and the other data set given next to the rows. Characteristics of the six ET data sets are provided in Table 1. Base from Natural Earth, 2024.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

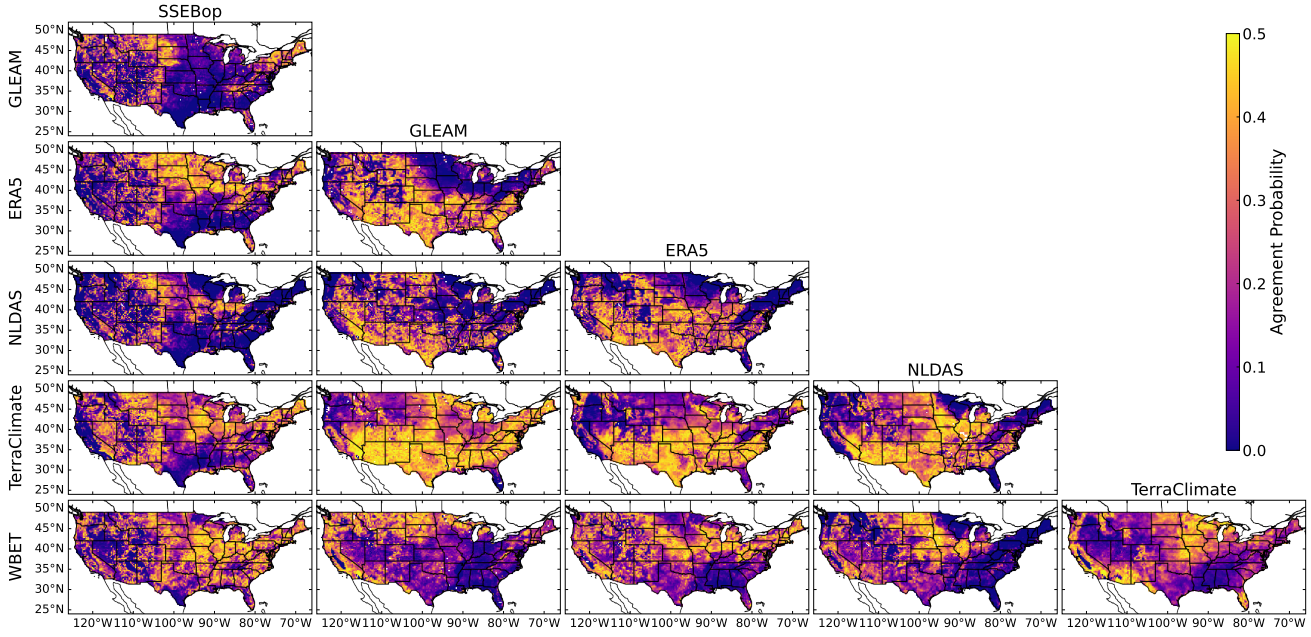

Figure A3. Median agreement probability grids of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair for the summer season. The figure panels are labeled with one data set in the pair being given above the columns and the other data set given next to the rows. Characteristics of the six ET data sets are provided in Table 1. Base from Natural Earth, 2024.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

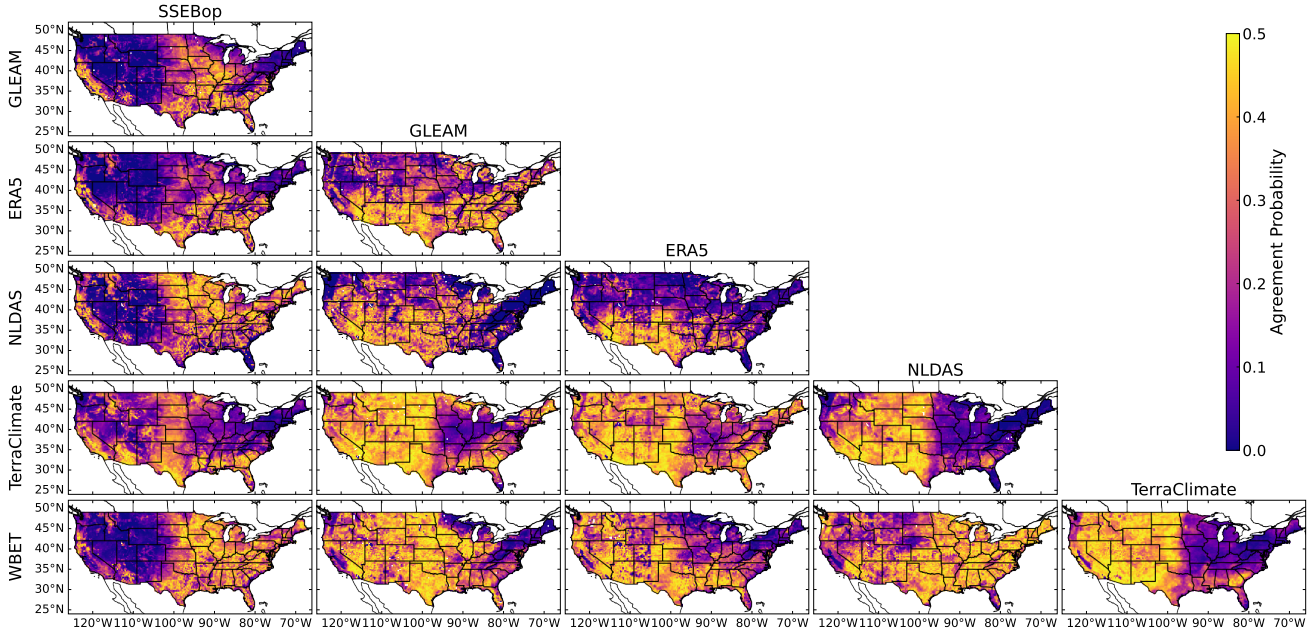

Figure A4. Median agreement probability grids of the six relative bias and extended collocation (EC) error covariance matrix estimates for each evapotranspiration (ET) data set pair for the fall season. The figure panels are labeled with one data set in the pair being given above the columns and the other data set given next to the rows. Characteristics of the six ET data sets are provided in Table 1. Base from Natural Earth, 2024.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Author contributions. K.D.: Conceptualization, Methodology, Analysis, Software Workflow, Writing—Original Draft; T.O.H.: Conceptualization, Writing; T.M.O.: Conceptualization, Methodology, Writing; and S.S.F.: Conceptualization, Project Administration, Writing

Competing interests. Authors K.D., T.O.H., T.M.O., and S.S.F. are employed by the U.S. Geological Survey and have no competing interests.

Disclaimer. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Acknowledgements. The authors sincerely thank Gabriel Senay and David Ketchum for providing valuable feedback on this manuscript. The work by Keith Doore was done while serving as a data scientist with the U.S. Geological Survey. Funding for this research was provided by the Hydro-terrestrial Earth Systems Testbed (HyTEST) project of the U.S. Geological Survey Integrated Water Prediction program.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

References

Abatzoglou, J., Dobrowski, S., Parks, S., and Hegewisch, K.: Monthly climate and climatic water balance for global terrestrial surfaces from 1958-2015, https://doi.org/10.7923/G43J3B0R, 2017.

Abatzoglou, J. T.: Development of gridded surface meteorological data for ecological applications and modelling, International Journal of Climatology, 33, 121–131, https://doi.org/10.1002/joc.3413, 2013.

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., and Hegewisch, K. C.: TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015, Scientific Data, 5, 170 191, https://doi.org/10.1038/sdata.2017.191, 2018.

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

Alexandris, S., ed.: Evapotranspiration - An Overview, InTech, ISBN 9789535111153, https://doi.org/10.5772/3383, 2013.

Anderson, M. C., Norman, J. M., Mecikalski, J. R., Otkin, J. A., and Kustas, W. P.: A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation, Journal of Geophysical Research: Atmospheres, 112, 2006JD007 506, https://doi.org/10.1029/2006JD007506, 2007.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., Seibert, J., Hakala, K., Bock, A., Wagener, T., Farmer, W. H., Andréassian, V., Attinger, S., Viglione, A., Knight, R., Markstrom, S., and Over, T.: Accelerating advances in continental domain hydrologic modeling, Water Resources Research, 51, 10 078–10 091, https://doi.org/10.1002/2015WR017498, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2015WR017498, 2015.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Allen, P. M.: CONTINENTAL SCALE SIMULATION OF THE HYDROLOGIC BALANCE [1], JAWRA Journal of the American Water Resources Association, 35, 1037–1051, https://doi.org/10.1111/j.1752-1688.1999.tb04192.x, 1999.

Baik, J., Park, J., Hao, Y., and Choi, M.: Integration of multiple drought indices using a triple collocation approach, Stochastic Environmental Research and Risk Assessment, 36, 1177–1195, https://doi.org/10.1007/s00477-021-02044-7, 2022.

Beven, K.: Rainfall-Runoff Modelling: The Primer, Wiley, 1 edn., ISBN 9780470714591 9781119951001, https://doi.org/10.1002/9781119951001, 2012.

Bhatt, R., Doelling, D., Haney, C., Scarino, B., and Gopalan, A.: Consideration of Radiometric Quantization Error in Satellite Sensor Cross-Calibration, Remote Sensing, 10, 1131, https://doi.org/10.3390/rs10071131, 2018.

Bosilovich, M. G., Chen, J., Robertson, F. R., and Adler, R. F.: Evaluation of Global Precipitation in Reanalyses, Journal of Applied Meteorology and Climatology, 47, 2279–2299, https://doi.org/10.1175/2008JAMC1921.1, 2008.

Caires, S. and Sterl, A.: Validation of ocean wind and wave data using triple collocation, Journal of Geophysical Research: Oceans, 108, 2002JC001 491, https://doi.org/10.1029/2002JC001491, 2003.

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

Chen, Z., Zhang, B., Zhang, H., and Zhang, W.: Vicarious Calibration of Beijing-1 Multispectral Imagers, Remote Sensing, 6, 1432–1450, https://doi.org/10.3390/rs6021432, 2014.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Choi, M., Kustas, W. P., Anderson, M. C., Allen, R. G., Li, F., and Kjaersgaard, J. H.: An intercomparison of three remote sensing-based surface energy balance algorithms over a corn and soybean production region (Iowa, U.S.) during SMACEX, Agricultural and Forest Meteorology, 149, 2082–2097, https://doi.org/10.1016/j.agrformet.2009.07.002, 2009.

Dennehy, K. F., Litke, D. W., and McMahon, P. B.: The High Plains Aquifer, USA: groundwater development and sustainability, Geological Society, London, Special Publications, 193, 99–119, https://doi.org/10.1144/GSL.SP.2002.193.01.09, 2002.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrology and Earth System Sciences, 13, 913–921, https://doi.org/10.5194/hess-13-913-2009, 2009.

Dobrowski, S. Z., Abatzoglou, J., Swanson, A. K., Greenberg, J. A., Mynsberge, A. R., Holden, Z. A., and Schwartz, M. K.: The climate velocity of the contiguous U nited S tates during the 20th century, Global Change Biology, 19, 241–251, https://doi.org/10.1111/gcb.12026, 2013.

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

Dong, J., Lei, F., and Wei, L.: Triple Collocation Based Multi-Source Precipitation Merging, Frontiers in Water, 2, 1, https://doi.org/10.3389/frwa.2020.00001, 2020a.

Dong, J., Wei, L., Chen, X., Duan, Z., and Lu, Y.: An instrument variable based algorithm for estimating cross-correlated hydrological remote sensing errors, Journal of Hydrology, 581, 124 413, https://doi.org/10.1016/j.jhydrol.2019.124413, 2020b.

Doore, K., Over, T., Hodson, T., and Foks, S.: Workflow Notebooks for Evaluating the Statistical Agreement between Gridded Evapotranspiration Data Sets in the Conterminous United States via Triple Collocation, Software release, https://doi.org/10.5066/P1VN9ENA, 2024.

Duan, Z., Duggan, E., Chen, C., Gao, H., Dong, J., and Liu, J.: Comparison of traditional method and triple collocation analysis for evaluation of multiple gridded precipitation products across Germany, Journal of Hydrometeorology, https://doi.org/10.1175/JHM-D-21-0049.1, 2021.

Fang, H., Wei, S., Jiang, C., and Scipal, K.: Theoretical uncertainty analysis of global MODIS, CYCLOPES, and GLOBCARBON LAI products using a triple collocation method, Remote Sensing of Environment, 124, 610–621, https://doi.org/10.1016/j.rse.2012.06.013, 2012.

Faunt, C. C.: Groundwater availability of the Central Valley Aquifer, California, US Geological Survey, https://doi.org/10.3133/pp1766, 2009.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, 37, 4302–4315, https://doi.org/10.1002/joc.5086, 2017.

Fisher, J. B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M. F., Hook, S., Baldocchi, D., Townsend, P. A., Kilic, A., Tu, K., Miralles, D. D., Perret, J., Lagouarde, J., Waliser, D., Purdy, A. J., French, A., Schimel, D., Famiglietti, J. S., Stephens, G., and Wood, E. F.: The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources, Water Resources Research, 53, 2618–2626, https://doi.org/10.1002/2016WR020175, 2017.

Fisher, J. B., Lee, B., Purdy, A. J., Halverson, G. H., Dohlen, M. B., Cawse-Nicholson, K., Wang, A., Anderson, R. G., Aragon, B., Arain, M. A., Baldocchi, D. D., Baker, J. M., Barral, H., Bernacchi, C. J., Bernhofer, C., Biraud, S. C., Bohrer, G., Brunsell, N., Cappelaere, B., Castro-Contreras, S., Chun, J., Conrad, B. J., Cremonese, E., Demarty, J., Desai, A. R., De Ligne, A., Foltýnová, L., Goulden, M. L., Griffis, T. J., Grünwald, T., Johnson, M. S., Kang, M., Kelbe, D., Kowalska, N., Lim, J., Maïnassara, I., McCabe, M. F., Missik, J. E., Mohanty, B. P., Moore, C. E., Morillas, L., Morrison, R., Munger, J. W., Posse, G., Richardson, A. D., Russell, E. S., Ryu, Y., Sanchez-Azofeifa, A., Schmidt, M., Schwartz, E., Sharp, I., Šigut, L., Tang, Y., Hulley, G., Anderson, M., Hain, C., French, A., Wood, E., and

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Hook, S.: ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station, Water Resources Research, 56, e2019WR026 058, https://doi.org/10.1029/2019WR026058, 2020.

Friedl, M., McIver, D., Hodges, J., Zhang, X., Muchoney, D., Strahler, A., Woodcock, C., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., and Schaaf, C.: Global land cover mapping from MODIS: algorithms and early results, Remote Sensing of Environment, 83, 287–302, https://doi.org/10.1016/S0034-4257(02)00078-0, 2002.

Fu, B.: On the calculation of the evaporation from land surface, Scientia Atmospherica Sinica, 5, 23, 1981.

Gruber, A., Su, C., Crow, W. T., Zwieback, S., Dorigo, W. A., and Wagner, W.: Estimating error cross-correlations in soil moisture data sets using extended collocation analysis, Journal of Geophysical Research: Atmospheres, 121, 1208–1219, https://doi.org/10.1002/2015JD024027, 2016a.

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

Gruber, A., Dorigo, W. A., Crow, W., and Wagner, W.: Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals, IEEE Transactions on Geoscience and Remote Sensing, 55, 6780–6792, https://doi.org/10.1109/TGRS.2017.2734070, 2017.

Gruber, A., De Lannoy, G., Albergel, C., Al-Yaari, A., Brocca, L., Calvet, J.-C., Colliander, A., Cosh, M., Crow, W., Dorigo, W., Draper, C., Hirschi, M., Kerr, Y., Konings, A., Lahoz, W., McColl, K., Montzka, C., Muñoz-Sabater, J., Peng, J., Reichle, R., Richaume, P., Rüdiger, C., Scanlon, T., Van Der Schalie, R., Wigneron, J.-P., and Wagner, W.: Validation practices for satellite soil moisture retrievals: What are (the) errors?, Remote Sensing of Environment, 244, 111 806, https://doi.org/10.1016/j.rse.2020.111806, 2020.

Guse, B., Fatichi, S., Gharari, S., and Melsen, L. A.: Advancing Process Representation in Hydrological Models: Integrating New Concepts, Knowledge, and Data, Water Resources Research, 57, e2021WR030 661, https://doi.org/10.1029/2021WR030661, 2021.

Hamon, W. R.: Estimating Potential Evapotranspiration, Journal of the Hydraulics Division, 87, 107–120, https://doi.org/10.1061/JYCEAJ.0000599, 1961.

Harris, I., Jones, P., Osborn, T., and Lister, D.: Updated high-resolution grids of monthly climatic observations – the CRU TS3 .10 Dataset, International Journal of Climatology, 34, 623–642, https://doi.org/10.1002/joc.3711, 2014.

He, S., Smirnova, T. G., and Benjamin, S. G.: Single-Column Validation of a Snow Subgrid Parameterization in the Rapid Update Cycle Land-Surface Model (RUC LSM), Water Resources Research, 57, e2021WR029 955, https://doi.org/10.1029/2021WR029955, 2021.

Jia, Y., Li, C., Yang, H., Yang, W., and Liu, Z.: Assessments of three evapotranspiration products over China using extended triple collocation and water balance methods, Journal of Hydrology, 614, 128 594, https://doi.org/10.1016/j.jhydrol.2022.128594, 2022.

Jiang, C., Ryu, Y., Fang, H., Myneni, R., Claverie, M., and Zhu, Z.: Inconsistencies of interannual variability and trends in long-term satellite leaf area index products, Global Change Biology, 23, 4133–4146, https://doi.org/10.1111/gcb.13787, 2017.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Scientific Data, 6, 74, https://doi.org/10.1038/s41597-019-0076-8, 2019.

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

Khan, M. S., Baik, J., and Choi, M.: Inter-comparison of evapotranspiration datasets over heterogeneous landscapes across Australia, Advances in Space Research, 66, 533–545, https://doi.org/10.1016/j.asr.2020.04.037, 2020.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Kim, H., Parinussa, R., Konings, A. G., Wagner, W., Cosh, M. H., Lakshmi, V., Zohaib, M., and Choi, M.: Global-scale assessment and combination of SMAP with ASCAT (active) and AMSR2 (passive) soil moisture products, Remote Sensing of Environment, 204, 260–275, https://doi.org/10.1016/j.rse.2017.10.026, 2018.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, Journal of the Meteorological Society of Japan. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Levin, S. B., Briggs, M. A., Foks, S. S., Goodling, P. J., Raffensperger, J. P., Rosenberry, D. O., Scholl, M. A., Tiedeman, C. R., and Webb, R. M.: Uncertainties in measuring and estimating water-budget components: Current state of the science, WIREs Water, 10, e1646, https://doi.org/10.1002/wat2.1646, 2023.

Li, C., Tang, G., and Hong, Y.: Cross-evaluation of ground-based, multi-satellite and reanalysis precipitation products: Applicability of the Triple Collocation method across Mainland China, Journal of Hydrology, 562, 71–83, https://doi.org/10.1016/j.jhydrol.2018.04.039, 2018.

Li, C., Yang, H., Yang, W., Liu, Z., Jia, Y., Li, S., and Yang, D.: Error characterization of global land evapotranspiration products: Collocation-based approach, Journal of Hydrology, 612, 128 102, https://doi.org/10.1016/j.jhydrol.2022.128102, 2022.

Li, X., Zhang, W., Vermeulen, A., Dong, J., and Duan, Z.: Triple collocation-based merging of multi-source gridded evapotranspiration data in the Nordic Region, Agricultural and Forest Meteorology, 335, 109 451, https://doi.org/10.1016/j.agrformet.2023.109451, 2023.

Liu, H., Xin, X., Su, Z., Zeng, Y., Lian, T., Li, L., Yu, S., and Zhang, H.: Intercomparison and evaluation of ten global ET products at site and basin scales, Journal of Hydrology, 617, 128 887, https://doi.org/10.1016/j.jhydrol.2022.128887, 2023.

Lyu, F., Tang, G., Behrangi, A., Wang, T., Tan, X., Ma, Z., and Xiong, W.: Precipitation Merging Based on the Triple Collocation Method Across Mainland China, IEEE Transactions on Geoscience and Remote Sensing, 59, 3161–3176, https://doi.org/10.1109/TGRS.2020.3008033, 2021.

Mallick, K., Wandera, L., Bhattarai, N., Hostache, R., Kleniewska, M., and Chormanski, J.: A Critical Evaluation on the Role of Aerodynamic and Canopy–Surface Conductance Parameterization in SEB and SVAT Models for Simulating Evapotranspiration: A Case Study in the Upper Biebrza National Park Wetland in Poland, Water, 10, 1753, https://doi.org/10.3390/w10121753, 2018.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geoscientific Model Development, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

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

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

McMillan, H. K., Westerberg, I. K., and Krueger, T.: Hydrological data uncertainty and its implications, WIREs Water, 5, e1319, https://doi.org/10.1002/wat2.1319, 2018.

Melton, F. S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J. B., Kilic, A., Senay, G. B., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., ReVelle, P. M., Schull, M., Wang, C., Yang, Y., and Anderson, R. G.: OpenET: Filling a Critical Data Gap

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

in Water Management for the Western United States, JAWRA Journal of the American Water Resources Association, 58, 971–994, https://doi.org/10.1111/1752-1688.12956, 2022.

Migliavacca, M., Musavi, T., Mahecha, M. D., Nelson, J. A., Knauer, J., Baldocchi, D. D., Perez-Priego, O., Christiansen, R., Peters, J., Anderson, K., Bahn, M., Black, T. A., Blanken, P. D., Bonal, D., Buchmann, N., Caldararu, S., Carrara, A., Carvalhais, N., Cescatti, A., Chen, J., Cleverly, J., Cremonese, E., Desai, A. R., El-Madany, T. S., Farella, M. M., Fernández-Martínez, M., Filippa, G., Forkel, M., Galvagno, M., Gomarasca, U., Gough, C. M., Göckede, M., Ibrom, A., Ikawa, H., Janssens, I. A., Jung, M., Kattge, J., Keenan, T. F., Knohl, A., Kobayashi, H., Kraemer, G., Law, B. E., Liddell, M. J., Ma, X., Mammarella, I., Martini, D., Macfarlane, C., Matteucci, G., Montagnani, L., Pabon-Moreno, D. E., Panigada, C., Papale, D., Pendall, E., Penuelas, J., Phillips, R. P., Reich, P. B., Rossini, M., Rotenberg, E., Scott, R. L., Stahl, C., Weber, U., Wohlfahrt, G., Wolf, S., Wright, I. J., Yakir, D., Zaehle, S., and Reichstein, M.: The three major axes of terrestrial ecosystem function, Nature, 598, 468–472, https://doi.org/10.1038/s41586-021-03939-9, 2021.

Miralles, D. G., Crow, W. T., and Cosh, M. H.: Estimating Spatial Sampling Errors in Coarse-Scale Soil Moisture Estimates Derived from Point-Scale Observations, Journal of Hydrometeorology, 11, 1423–1429, https://doi.org/10.1175/2010JHM1285.1, 2010.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453–469, https://doi.org/10.5194/hess-15-453-2011, 2011.

Montoya, R. D. and Osorio, A. F.: Methodology to Correct Wind Speed during Average Wind Conditions: Application to the Caribbean Sea, Journal of Atmospheric and Oceanic Technology, 31, 1922–1945, https://doi.org/10.1175/JTECH-D-13-00124.1, 2014.

Mu, Q., Heinsch, F. A., Zhao, M., and Running, S. W.: Development of a global evapotranspiration algorithm based on MODIS and global meteorology data, Remote Sensing of Environment, 111, 519–536, https://doi.org/10.1016/j.rse.2007.04.015, 2007.

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

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Muñoz Sabater, J.: ERA5-Land monthly averaged data from 1950 to present, https://doi.org/10.24381/CDS.68D2BB30, 2019.

Nearing, G. S., Yatheendradas, S., Crow, W. T., Bosch, D. D., Cosh, M. H., Goodrich, D. C., Seyfried, M. S., and Starks, P. J.: Nonparametric triple collocation, Water Resources Research, 53, 5516–5530, https://doi.org/10.1002/2017WR020359, 2017.

NLDAS Project: NLDAS Noah Land Surface Model L4 Monthly 0.125 x 0.125 degree V2.0, https://doi.org/10.5067/WB224IA3PVOJ, 2020.

Ochege, F. U., Shi, H., Li, C., Ma, X., Igboeli, E. E., and Luo, G.: Assessing Satellite, Land Surface Model and Reanalysis Evapotranspiration Products in the Absence of In-Situ in Central Asia, Remote Sensing, 13, 5148, https://doi.org/10.3390/rs13245148, 2021.

Oki, T. and Kanae, S.: Global Hydrological Cycles and World Water Resources, Science, 313, 1068–1072, https://doi.org/10.1126/science.1128845, 2006.

Park, J., Baik, J., and Choi, M.: Triple collocation-based multi-source evaporation and transpiration merging, Agricultural and Forest Meteorology, 331, 109 353, https://doi.org/10.1016/j.agrformet.2023.109353, 2023.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron,

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. d., Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. d., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A. H., Gough, C. M., Goulden, M. L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B. U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W. L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A. J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H. A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J. H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A. M. S., Merbold, L., Meyer, W., Meyers, T., Miller, S. D., Minerbi, S., Moderow, U., Monson, R. K., Montagnani, L., Moore, C. E., Moors, E., Moreaux, V., Moureaux, C., Munger, J. W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J. E., Ourcival, J.-M., Papuga, S. A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R. P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S. M., Rambal, S., Rannik, U., Raz-Yaseef, N., Rebmann, C., Reed, D., Dios, V. R. d., Restrepo-Coupe, N., Reverter, B. R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S. R., Sánchez-Cañete, E. P., Sanchez-Mejia, Z. M., Schmid, H. P., Schmidt, M., Schneider, K., Schrader, F., Schroder, I., Scott, R. L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R. M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J. P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., and Papale, D.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Scientific Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Pierdicca, N., Fascetti, F., Pulvirenti, L., and Crapolicchio, R.: Error Characterization of Soil Moisture Satellite Products: Retrieving Error Cross-Correlation Through Extended Quadruple Collocation, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 4522–4530, https://doi.org/10.1109/JSTARS.2017.2714025, 2017.

Priestley, C. H. B. and Taylor, R. J.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, Monthly Weather Review, 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2, 1972.

Reichle, R. H., Draper, C. S., Liu, Q., Girotto, M., Mahanama, S. P. P., Koster, R. D., and De Lannoy, G. J. M.: Assessment of MERRA-2 Land Surface Hydrology Estimates, Journal of Climate, 30, 2937–2960, https://doi.org/10.1175/JCLI-D-16-0720.1, 2017.

Reitz, M., Senay, G., and Sanford, W.: Combining Remote Sensing and Water-Balance Evapotranspiration Estimates for the Conterminous United States, Remote Sensing, 9, 1181, https://doi.org/10.3390/rs9121181, 2017.

Reitz, M., Sanford, W. E., and Saxe, S.: Ensemble Estimation of Historical Evapotranspiration for the Conterminous U.S., Water Resources Research, 59, e2022WR034 012, https://doi.org/10.1029/2022WR034012, 2023a.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Reitz, M. M., Sanford, W. E., and Saxe, S. W.: Historical Evapotranspiration for the Conterminous U.S., https://doi.org/10.5066/P9EZ3VAS, 2023b.

Ribal, A. and Young, I. R.: Global Calibration and Error Estimation of Altimeter, Scatterometer, and Radiometer Wind Speed Using Triple Collocation, Remote Sensing, 12, 1997, https://doi.org/10.3390/rs12121997, 2020.

Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G., William Munger, J., Ricciuto, D. M., Stoy, P. C., Suyker, A. E., Verma, S. B., and Wofsy, S. C.: A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes, Agricultural and Forest Meteorology, 136, 1–18, https://doi.org/10.1016/j.agrformet.2006.01.007, 2006.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bulletin of the American Meteorological Society, 85, 381–394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

Roebeling, R. A., Wolters, E. L. A., Meirink, J. F., and Leijnse, H.: Triple Collocation of Summer Precipitation Retrievals from SEVIRI over Europe with Gridded Rain Gauge and Weather Radar Data, Journal of Hydrometeorology, 13, 1552–1566, https://doi.org/10.1175/JHM-D-11-089.1, 2012.

Saxe, S., Farmer, W., Driscoll, J., and Hogue, T. S.: Implications of model selection: a comparison of publicly available, conterminous US-extent hydrologic component estimates, Hydrology and Earth System Sciences, 25, 1529–1568, https://doi.org/10.5194/hess-25-1529-2021, 2021.

Scipal, K., Holmes, T., De Jeu, R., Naeimi, V., and Wagner, W.: A possible solution for the problem of estimating the error structure of global soil moisture data sets, Geophysical Research Letters, 35, 2008GL035 599, https://doi.org/10.1029/2008GL035599, 2008.

Senay, G. and Kagone, S.: Daily SSEBop Evapotranspiration Data from 2000 to 2018, https://doi.org/10.5066/P9L2YMV, 2019.

Senay, G. B.: Satellite Psychrometric Formulation of the Operational Simplified Surface Energy Balance (SSEBop) Model for Quantifying and Mapping Evapotranspiration, Applied Engineering in Agriculture, 34, 555–566, https://doi.org/10.13031/aea.12614, 2018.

Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H., and Verdin, J. P.: Operational Evapotranspiration Mapping Using Remote Sensing and Weather Datasets: A New Parameterization for the SSEB Approach, JAWRA Journal of the American Water Resources Association, 49, 577–591, https://doi.org/10.1111/jawr.12057, 2013.

Singh, V. P., Singh, R., Paul, P. K., Bisht, D. S., and Gaur, S.: Uncertainty Analysis in Hydrologic Modelling, in: Hydrological Processes Modelling and Data Analysis, vol. 127, pp. 203–227, Springer Nature Singapore, Singapore, ISBN 9789819713158 9789819713165, https://doi.org/10.1007/978-981-97-1316-5_10, 2024.

Siqueira, V. A., Paiva, R. C. D., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., Paris, A., Calmant, S., and Collischonn, W.: Toward continental hydrologic–hydrodynamic modeling in South America, Hydrology and Earth System Sciences, 22, 4815–4842, https://doi.org/10.5194/hess-22-4815-2018, 2018.

Stanton, J. S., Qi, S. L., Ryter, D. W., Falk, S. E., Houston, N. A., Peterson, S. M., Westenbroek, S. M., and Christenson, S. C.: Selected approaches to estimate water-budget components of the High Plains, 1940 through 1949 and 2000 through 2009, US Geological Survey, https://doi.org/10.3133/sir20115183, 2011.

Stoffelen, A.: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, Journal of Geophysical Research: Oceans, 103, 7755–7766, https://doi.org/10.1029/97JC03180, 1998.

Su, C., Ryu, D., Crow, W. T., and Western, A. W.: Beyond triple collocation: Applications to soil moisture monitoring, Journal of Geophysical Research: Atmospheres, 119, 6419–6439, https://doi.org/10.1002/2013JD021043, 2014.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z., and Hong, Y.: Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets, Remote Sensing of Environment, 240, 111 697, https://doi.org/10.1016/j.rse.2020.111697, 2020.

Tang, R. and Li, Z.: Estimating Daily Evapotranspiration From Remotely Sensed Instantaneous Observations With Simplified Derivations of a Theoretical Model, Journal of Geophysical Research: Atmospheres, 122, https://doi.org/10.1002/2017JD027094, 2017.

Tillman, F. D., Day, N. K., Miller, M. P., Miller, O. L., Rumsey, C. A., Wise, D. R., Longley, P. C., and McDonnell, M. C.: A Review of Current Capabilities and Science Gaps in Water Supply Data, Modeling, and Trends for Water Availability Assessments in the Upper Colorado River Basin, Water, 14, 3813, https://doi.org/10.3390/w14233813, 2022.

Torres-Rojas, L., Vergopolan, N., Herman, J. D., and Chaney, N. W.: Towards an Optimal Representation of Sub-Grid Heterogeneity in Land Surface Models, Water Resources Research, 58, e2022WR032 233, https://doi.org/10.1029/2022WR032233, 2022.

Trebs, I., Mallick, K., Bhattarai, N., Sulis, M., Cleverly, J., Woodgate, W., Silberstein, R., Hinko-Najera, N., Beringer, J., Meyer, W. S., Su, Z., and Boulet, G.: The role of aerodynamic resistance in thermal remote sensing-based evapotranspiration models, Remote Sensing of Environment, 264, 112 602, https://doi.org/10.1016/j.rse.2021.112602, 2021.

U.S. Geological Survey: Spatial Provinces and Domains of the Central Valley for Textural Analysis, https://doi.org/10.5066/P9Y4MR92, 2023a.

U.S. Geological Survey: Digital map of aquifer boundary for the High Plains aquifer in parts of Colorado, Kansas, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wyoming, https://doi.org/10.5066/P9KA17IW, 2023b.

U.S. Geological Survey: Upper Colorado River Basin Boundary - ScienceBase-Catalog, https://www.sciencebase.gov/catalog/item/4f4e4a38e4b07f02db61cebb, 2024.

Van Dijk, A. I. J. M., Renzullo, L. J., Wada, Y., and Tregoning, P.: A global water cycle reanalysis (2003–2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble, Hydrology and Earth System Sciences, 18, 2955–2973, https://doi.org/10.5194/hess-18-2955-2014, 2014.

Volk, J. M., Huntington, J., Melton, F. S., Allen, R., Anderson, M. C., Fisher, J. B., Kilic, A., Senay, G., Halverson, G., Knipper, K., Minor, B., Pearson, C., Wang, T., Yang, Y., Evett, S., French, A. N., Jasoni, R., and Kustas, W.: Development of a Benchmark Eddy Flux Evapotranspiration Dataset for Evaluation of Satellite-Driven Evapotranspiration Models Over the CONUS, Agricultural and Forest Meteorology, 331, 109 307, https://doi.org/10.1016/j.agrformet.2023.109307, 2023.

Volk, J. M., Huntington, J. L., Melton, F. S., Allen, R., Anderson, M., Fisher, J. B., Kilic, A., Ruhoff, A., Senay, G. B., Minor, B., Morton, C., Ott, T., Johnson, L., Comini De Andrade, B., Carrara, W., Doherty, C. T., Dunkerly, C., Friedrichs, M., Guzman, A., Hain, C., Halverson, G., Kang, Y., Knipper, K., Laipelt, L., Ortega-Salazar, S., Pearson, C., Parrish, G. E. L., Purdy, A., ReVelle, P., Wang, T., and Yang, Y.: Assessing the accuracy of OpenET satellite-based evapotranspiration data to support water resource and land management applications, Nature Water, 2, 193–205, https://doi.org/10.1038/s44221-023-00181-7, 2024.

Wang-Erlandsson, L., Bastiaanssen, W. G. M., Gao, H., Jägermeyr, J., Senay, G. B., van Dijk, A. I. J. M., Guerschman, J. P., Keys, P. W., Gordon, L. J., and Savenije, H. H. G.: Global root zone storage capacity from satellite-based evaporation, Hydrology and Earth System Sciences, 20, 1459–1481, https://doi.org/10.5194/hess-20-1459-2016, 2016.

Waterman, T., Bragg, A. D., Hay-Chapman, F., Dirmeyer, P. A., Fowler, M. D., Simon, J., and Chaney, N.: A Two-Column Model Parameterization for Subgrid Surface Heterogeneity Driven Circulations, Journal of Advances in Modeling Earth Systems, 16, e2023MS003 936, https://doi.org/10.1029/2023MS003936, 2024.

https://doi.org/10.5194/egusphere-2026-2036
Preprint. Discussion started: 29 April 2026

Willmott, C. J. and Robeson, S. M.: Climatologically aided interpolation (CAI) of terrestrial air temperature, International Journal of Climatology, 15, 221–229, https://doi.org/10.1002/joc.3370150207, 1995.

Willmott, C. J., Rowe, C. M., and Mintz, Y.: Climatology of the terrestrial seasonal water cycle, Journal of Climatology, 5, 589–606, https://doi.org/10.1002/joc.3370050602, 1985.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, Journal of Geophysical Research: Atmospheres, 117, 2011JD016 048, https://doi.org/10.1029/2011JD016048, 2012.

Xu, L., Chen, N., Zhang, X., Moradkhani, H., Zhang, C., and Hu, C.: In-situ and triple-collocation based evaluations of eight global root zone soil moisture products, Remote Sensing of Environment, 254, 112 248, https://doi.org/10.1016/j.rse.2020.112248, 2021.

Yang, Y., Chen, R., Han, C., and Qing, W.: Measurement and estimation of the summertime daily evapotranspiration on alpine meadow in the Qilian Mountains, northwest China, Environmental Earth Sciences, https://doi.org/10.1007/s12665-012-1907-5, 2012.

Yilmaz, M. T. and Crow, W. T.: Evaluation of Assumptions in Soil Moisture Triple Collocation Analysis, Journal of Hydrometeorology, 15, 1293–1302, https://doi.org/10.1175/JHM-D-13-0158.1, 2014.

Yilmaz, M. T., Crow, W. T., Anderson, M. C., and Hain, C.: An objective methodology for merging satellite- and model-based soil moisture products, Water Resources Research, 48, 2011WR011 682, https://doi.org/10.1029/2011WR011682, 2012.

Yin, G. and Park, J.: The use of triple collocation approach to merge satellite- and model-based terrestrial water storage for flood potential analysis, Journal of Hydrology, 603, 127 197, https://doi.org/10.1016/j.jhydrol.2021.127197, 2021.

Zhang, L., Hickel, K., Dawes, W. R., Chiew, F. H. S., Western, A. W., and Briggs, P. R.: A rational function approach for estimating mean annual evapotranspiration, Water Resources Research, 40, 2003WR002 710, https://doi.org/10.1029/2003WR002710, 2004.

Zhang, Y., Kong, D., Gan, R., Chiew, F. H., McVicar, T. R., Zhang, Q., and Yang, Y.: Coupled estimation of 500 m and 8-day resolution global evapotranspiration and gross primary production in 2002–2017, Remote Sensing of Environment, 222, 165–182, https://doi.org/10.1016/j.rse.2018.12.031, 2019.

Zheng, Z., Ma, Z., Li, M., and Xia, J.: Regional water budgets and hydroclimatic trend variations in Xinjiang from 1951 to 2000, Climatic Change, 144, 447–460, https://doi.org/10.1007/s10584-016-1842-7, 2017.

Zhou, S., Wang, Y., Li, Z., Chang, J., and Guo, A.: Quantifying the Uncertainty Interaction Between the Model Input and Structure on Hydrological Processes, Water Resources Management, 35, 3915–3935, https://doi.org/10.1007/s11269-021-02883-7, 2021.

Zhu, B., Xie, X., Wang, Y., and Zhao, X.: The Benefits of Continental-Scale High-Resolution Hydrological Modeling in the Detection of Extreme Hydrological Events in China, Remote Sensing, 15, 2402, https://doi.org/10.3390/rs15092402, 2023.

Zwieback, S., Scipal, K., Dorigo, W., and Wagner, W.: Structural and statistical properties of the collocation technique for error characterization, Nonlinear Processes in Geophysics, 19, 69–80, https://doi.org/10.5194/npg-19-69-2012, 2012.