Leveraging pre-storm soil moisture estimates for enhanced land surface model calibration in ungauged hydrologic basins

Wade T. Crow[1], Jianzhi Dong[1,2] and Rolf H. Reichle[3]

[1] USDA Hydrology and Remote Sensing Laboratory, Beltsville, MD
[2] Massachusetts Institute of Technology, Cambridge, MA
[3] Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD

Submitted to *Water Resources Research*

Abstract

Despite long-standing efforts, hydrologists still lack robust tools for calibrating land surface model (LSM) streamflow estimates within ungauged basins. Using surface soil moisture estimates from the Soil Moisture Active Passive Level 4 Soil Moisture (L4_SM) product, precipitation observations, and streamflow gauge measurements for 617 medium-scale (200-10,000 km^2) basins in the contiguous United States, we measure the temporal (Spearman) rank correlation between antecedent (i.e., pre-storm) surface soil moisture (ASM) and the storm-scale runoff coefficient (RC; the fraction of storm-scale precipitation accumulation converted into streamflow). In humid and semi-humid basins, this rank correlation is shown to be sufficiently strong to allow for the substitution of storm-scale RC observations (available only in basins that are both lightly regulated and gauged) with high-quality ASM values (available quasi-globally from L4_SM) in streamflow calibration procedures. Using this principle, we define a new, basin-wise LSM streamflow calibration approach based on L4_SM alone and successfully apply it to identify LSM configurations that produce a high rank correlation with observed RC. However, since the approach cannot detect RC bias, it is less successful in identifying LSM configurations with low mean-absolute error.

Plain Text Summary

Accurately forecasting the fraction of rainfall that runs off into streams, as opposed to infiltrates into the soil, is critical for flash-flood prediction, water-resource monitoring, and tracking the transport of nutrients from agricultural fields into local waterways. Such forecasting is typically performed by hydrologic models that attempt to represent the physical processes responsible for surface runoff generation. However, to provide accurate streamflow forecasts, these models typically need to be calibrated against actual streamflow observations. This is problematic given the relatively poor, and declining, global availability of stream gauges. This paper presents a novel model calibration strategy that uses soil moisture from remote sensing and numerical modeling in place of streamflow observations during calibration. This transition has significant practical advantages because, unlike streamflow observations, the soil moisture data are continuously available across space. Our results demonstrate that this new approach can significantly improve hydrologic models within humid and semi-humid basins lacking sufficient ground-based instrumentation for traditional streamflow calibration.

1. Introduction

Despite several decades of development, land surface models (LSMs) still do not generally provide adequate streamflow estimates outside of hydrologic basins in which they have been directly calibrated (Xia et al., 2012; Hrachowitz et al., 2013). This is problematic due to the limited, and declining, worldwide availability of streamflow gauge data (Fekete et al., 2015) – as well as the proliferation of stream diversion and impoundment infrastructure that degrades the quality of hydrologic information contained in streamflow observations. As a result, there is a widely acknowledged need to develop effective LSM calibration strategies that can be applied in the absence of reliable streamflow observations (Samaniego et al., 2017).

The estimation and routing of runoff is a multi-faceted process; however, one fundamental aspect is the application of an LSM to estimate the fraction of storm-scale precipitation converted into runoff (hereinafter, the runoff coefficient or "RC"). Storm-to-storm variation in antecedent (i.e., pre-storm) soil moisture (ASM) is a well-known predictor of RC - see, e.g., Song & Wang (2019) and references therein. However, past attempts to quantify the impact of ASM on RC have been complicated by the presence of significant independent errors in available ASM and RC estimates and the resulting attenuation bias in their sampled temporal correlation (Crow et al., 2017). Attenuation bias refers to the tendency for independent random errors, present in either independent or dependent variables, to spuriously decrease sampled cross-correlation values (Hutcheon et al., 2010). Recent work with the Soil Moisture Active Passive (SMAP) Level 4 Surface and Root-zone Soil Moisture (L4_SM) product suggests that, once attenuation bias is minimized via the application of high-quality L4_SM ASM estimates, storm-to-storm variations in ASM can be shown to play a dominant role in driving RC temporal variability within the central and eastern United States (Crow et al., 2019). This result stands in contrast with the typical representation of RC in LSMs - which generally predict a weaker role for ASM in determining RC (Crow et al., 2018; 2019).

The apparent strength of the true coupling between ASM and RC presents an opportunity for LSM streamflow calibration. Crow et al. (2019) suggest that, in certain cases, the relationship between ASM estimates acquired from the L4_SM product and storm-scale RC observations is sufficiently strong for the two quantities to be used inter-changeably in correlation-based calibration objective functions. This is notable because the L4_SM product is available globally - while meaningful RC observations are restricted to a relatively small number of lightly regulated hydrologic basins with streamflow measurements available at their outlet. Therefore, if L4_SM ASM and gauge-based RC values can be applied interchangeably, new opportunities exist for expanding the meaningful calibration of LSMs into ungauged or highly regulated basins – at least in areas like the contiguous United States (CONUS) where the L4_SM product is known to provide high-quality ASM information (Crow et al., 2017).

Hereinafter, we will refer to the assumption of perfect temporal (i.e., storm-to-storm) rank correlation between ASM and RC as the "perfect correlation" (PC) assumption. Our two key objectives are to: i) evaluate the strength of the PC assumption using observations and ii) investigate the potential of the PC assumption as a streamflow calibration principle for LSMs.

To achieve our first objective, we use ASM estimates acquired from the L4_SM product (Section 2.2) and RC estimates based on streamflow and rainfall observations (Section 2.3) obtained within the set of lightly regulated basins described in Section 2.1. Key steps towards achieving this objective include the discrimination of individual storm events (Section 3.1), modeling the impact of random errors on estimated ASM and RC values (Section 3.2), and the identification of factors impacting the observed correlation between ASM and RC (Section 2.4). Results regarding the strength of the PC assumption are then described in Section 5.1.

Our second objective expands on the first by evaluating LSM calibration strategies based on the PC assumption. These strategies employ an ensemble of LSM model configurations (Sections 4.1-4.3) and are described in detail in Section 4.4. Evaluation metrics, used to assess the performance of various calibration strategies, are introduced in Section 4.5. Note that, in this context, the term "calibration" indicates the selection of an optimal LSM ensemble member for an individual basin based on a particular calibration strategy. LSM calibration results based on the PC assumption are then described in Sections 5.2-5.4.

2. Domain and Data

2.1. Study basins

Our analysis is based on an examination of ASM, precipitation, and streamflow data within 617 medium-scale, lightly regulated CONUS basins during the period 1 April 2015 to 30 August 2020. Here, we provide background on our study basins and the datasets used to examine the relationship between ASM and RC.

Study basins are based on a list of 1145 lightly regulated CONUS basins described in Lohmann et al. (2004) and examined in Xia et al. (2012). From this original list, only basins between 200 and 10,000 km^2 in size are considered here. Likewise, basins containing fewer than 50 snow-free and frozen-soil-free storm events during our analysis period are discarded - see additional details on our applied storm-event definition in Section 3.1. Finally, a small number (< 10) of additional basins containing visible reservoirs (despite prior screening for regulation), suffering from significant temporal gaps in daily USGS streamflow observations, or providing clearly non-physical long-term streamflow statistics (e.g., mean-annual streamflow exceeding mean-annual precipitation) are removed. Such screening results in the 617 selected basins shown in Figure 1. These basins are generally restricted to the eastern half of CONUS – along with a smaller number of basins along the west coast.

Daily USGS streamflow observations are acquired and processed for each basin outlet. Note that, in the interest of maximizing the spatial coverage of our analysis, the rain-gauge density threshold suggested by Schaake et al. (2000) is not applied. Therefore, significant precipitation measurement errors are still possible. Likewise, despite our best efforts, the absence of small-scale anthropogenic impoundment or diversion structures cannot be guaranteed.

2.2 SMAP L4_SM product

ASM values are based on the area-weighted spatial averaging of 9-km resolution surface (0-5 cm) soil moisture values acquired from Version 5 of the L4_SM product (Reichle et al., 2019; 2020; 2021a) within each study basin. The 3-hourly L4_SM product is generated through the sequential assimilation of SMAP brightness temperature data (Piepmeier et al., 2017) into the NASA Catchment LSM (Reichle et al., 2017). Hourly, 0.25-degree surface meteorological forcing data for the Catchment LSM is derived from the Goddard Earth Observing System Forward-Processing (GEOS-FP) product (https://gmao.gsfc.nasa.gov/GMAO_products/; Lucchesi, 2018). Over CONUS, the GEOS-FP precipitation forcing is corrected to match daily accumulations provided by the gauge-based NOAA Climate Prediction Center Unified (CPCU) product at a 0.5-degree scale. Prior to the start of our analysis (1 April 2015), the Catchment LSM is spun up from a cold start on 1 January 1980 using forcing data acquired from the Modern-Era Retrospective Analysis for Research and Applications, Version-2 dataset (Gelaro et al., 2017).

Daily ASM values are based on L4_SM surface soil moisture estimates within the 3-hour time window (centered at 2230 UTC) closest to the end of each 0 to 24 UTC day. Sampling at the end of the UTC day ensures that such ASM values are acquired as close as possible to the potential start of a rainfall event on the following day.

Prior work has established that the L4_SM product provides a significantly better pre-storm indicator of ASM than other available SM products - including a model-only version of the L4_SM product that does not assimilate SMAP brightness temperature (Crow et al., 2017). Likewise, L4_SM pre-storm surface (0 to 5-cm) soil moisture estimates are used because past work suggests that they correlate slightly better with basin runoff response than corresponding root-zone (0 to 1-m) L4_SM estimates (Crow et al., 2018).

2.3 NLDAS-2/CPCU precipitation

Daily (i.e., 0 to 24 UTC) precipitation totals are based on the spatial weighted averaging of 0.125-degree gridded NLDAS-2 precipitation (Xia et al., 2009) estimates falling within each basin. At a daily scale, these estimates are designed to match a 0.25-degree gauge-only CPCU product (Chen et al., 2008) and corrected for topographic effects by the Parameter-elevation Regressions on Independent Slopes Mode precipitation climatology dataset (Daly et al., 1994). Note that we use the NLDAS-2 forcing data, instead of the GEOS-FP precipitation data applied to force the land model in the L4_SM system, to maximize the independence of ASM and RC estimates used in our analysis of the PC assumption. Nevertheless, there is considerable overlap in the rain gauge data used as the basis for both. The potential impact of cross-dependency in the precipitation data used to generate both ASM and RC is discussed further in Section 3.2.

2.4 Aridity index

Aridity index (AI) values, applied below to explain spatial trends in observed ASM versus RC coupling, are taken from the Global Aridity and Potential Evaporation Dataset (Zomer et al., 2007; Zomer et al., 2008). This product conforms to the AI definition offered by the United Nations Environment Program (UNEP, 1992) whereby AI is the dimensionless ratio between long-term, mean-annual precipitation divided by long-term, mean-annual potential evapotranspiration (i.e., low AI values correspond to arid climates and high AI values to humid climates).

AI values based on this definition are typically bounded between 0 and 1.5. Any absolute labelling of AI values is somewhat subjective; however, a value of 0.4 approximates the well-known wet/dry climate transition along the 100th meridian in CONUS. Note that AI values are based on climatological averages sampled from long-term (1950 to 2000) observations and an implied assumption of climate stationarity. As such, their temporal support does not correspond directly to our 1 April 2015 to 30 August 2020 analysis period.

3. Approach

3.1 Storm events and SRCS

Storm-event separation is based on the approach of Crow et al. (2017) where a new storm event is assumed to start on any day with a NLDAS-2/CPCU precipitation accumulation exceeding P_{min}. Unless otherwise noted, P_{min} = 10 mm d^{-1}.

Following a triggering daily rainfall accumulation, storm events are assumed to last for an N-day period defined by rounding up the basin saturation time expression of Linsley et al. (1982) to the nearest positive integer:

$$N \text{ [days]} = \mathrm{CEIL}[(A * 2.59)^{0.2}] \quad (1)$$

where A is basin area [km^2] and CEIL is an upward integer rounding function. As a result, N is meant to capture the period in which streamflow can be attributed to a given storm event. Here, it is assumed to be independent of rainfall-event size. Derived values of N range from 4 days for our smallest (~200 km^2) basin to 8 days for our largest (~10,000 km^2) basin.

ASM refers to the lowest end-of-day (i.e., 24 UTC or the closest available alternative), basin-averaged L4_SM surface soil moisture (0-5 cm) value for the two-day period preceding the start of a storm event, as will be discussed further below. Storm events interrupted by a new storm (i.e., the arrival of another daily precipitation accumulation exceeding $P_{\min}$ within an earlier N-day storm event) are discarded, and a new event is assumed to begin coincident with the latest triggering precipitation event.

Since our analysis focuses only on hydrological responses to rainfall incident on unfrozen and non-snow-covered soil, only storm events where the pre-storm, basin-averaged snow fractional cover is below 1% (by area) and the 24 UTC surface temperature is above 3° C on the first (i.e., triggering) day of the event are considered. Surface state results from a baseline LSM configuration (see Section 4 below) are used for assessing both thresholds – which mask snow, rain-on-snow, and rain-on-frozen-soil storm events from our analysis.

As noted above, storm-scale RC is calculated as the streamflow volume during the storm event divided by total precipitation accumulation volume (during the same storm-event sampling period). Within basins containing more than 50 valid storm events during the study period (1 April 2015 to 30 August 2020), soil moisture runoff coupling strength (SRCS) is defined as the sampled Spearman rank correlation between ASM and RC values across all storm events. A minimum of 50 qualifying events is required, reflecting a trade-off between the competing concerns of maximizing the spatial coverage of our analysis versus adequately filtering basins where SRCS values cannot be accurately sampled.

As discussed above, the perfect rank correlation (PC) assumption dictates that SRCS = 1. That is, it assumes that rank variations in RC across multiple storm events can be captured perfectly given appropriate knowledge of ASM. As a correlation-based metric, SRCS is not impacted by the potential presence of bias in the L4_SM product.

The separation of a continuous rainfall time series into discrete multi-day events introduces potential ambiguities into the calculation of SRCS. One potentially problematic case is when rainfall begins before 24 UTC on a given day (without exceeding the $P_{\min}$ daily accumulation threshold) and then continues into the following day – when the daily $P_{\min}$ accumulation threshold is exceeded and, as a result, a new storm event begins. In this case, "end-of-day" surface soil moisture (SSM) may be enhanced on the day prior to the event, because rainfall actually started before 24 UTC, and storm-scale RC will also be spuriously increased, since rainfall during the previous day will be neglected in the "storm-scale" calculation of RC. Such simultaneous enhancement to both ASM and RC could, conceivably, inflate sampled SRCS values. To combat this, we define ASM as the minimum end-of-day SSM for the two-day period prior to the start of a storm event. Therefore, in the case where ASM is spuriously inflated by an event starting before 24 UTC, ASM will be defined using end-of-day SSM for the *previous* day - and thereby avoid any spurious increase in ASM.

This approach has the benefit of not discarding any qualifying storm events. A more conservative approach is to simply discard events that are preceded by more than trace amounts of daily precipitation (defined here as any daily accumulation exceeding $P_{\min}/5$). While this causes a significant reduction in the number of storm events available for sampling SRCS, it also provides an important check that our SRCS results are not being spuriously impacted by rainfall events crossing over the 24 UTC demarcation. Therefore, key results below will be re-generated using this more stringent masking procedure to ensure that our main SRCS results are reliable (see Section 5.1 and Figure S.1 of the Supporting Information).

3.2 Error models for ASM and RC estimates

As described above, a key focus of our analysis is the absolute value of SRCS sampled from noisy ASM and RC estimates. As a correlation-based metric, SRCS will be spuriously biased towards zero (i.e., attenuated) via the presence of independent, random error in L4_SM ASM and/or USGS/CPCU RC estimates. This attenuation bias is not associated with sample size limitations and will persist even in the theoretical case of an infinite sample size (Dong and Crow, 2018). It is, therefore, conceptually distinct from the representation of confidence intervals that vary as a function of sample size.

To estimate attenuation bias in sampled SRCS values, we first define random error models for the L4_SM ASM analysis and the NLDAS-2/CPCU precipitation observations underlying the RC estimates. L4_SM estimates are assumed to be degraded by mean-zero, time-independent, additive Gaussian random error with a standard deviation of σ_{L4_SM} = 0.032 m^3m^{-3}. This estimate is based on L4_SM ground validation results against in situ measurements at the 36-km scale (Reichle et al., 2017) and a minor adjustment to correct for random uncertainty in the ground measurements themselves (Chen et al., 2019). While there is almost certainly spatial and temporal variability in σ_{L4_SM} (Qiu et al., 2021), such variability is difficult to assess and accurately reflect in an error model and therefore neglected here

Likewise, NLDAS-2/CPCU daily precipitation observations are assumed to be impacted by random, multiplicative errors sampled from a log-normal distribution with unit mean and standard deviation σ_{CPCU}. Here, σ_{CPCU} is estimated as a discrete function of the time-average number of CPCU rain gauges (N_G) contained within each of our study basins using the relative-accuracy versus gauge-density relationship summarized by Villarini et al. (2008). Specifically, Figure 9 in Villarini et al. (2008) suggests that:

$$\sigma_{CPCU} = 0.10 \quad \text{for } N_G \geq 9$$
$$\sigma_{CPCU} = 0.15 \quad \text{for } 5 \leq N_G \leq 8$$
$$\sigma_{CPCU} = 0.20 \quad \text{for } N_G = 4 \qquad (3)$$

$$\sigma_{CPCU} = 0.25 \text{ for } N_G = 3$$

$$\sigma_{CPCU} = 0.30 \text{ for } N_G = 2$$

$$\sigma_{CPCU} = 0.40 \text{ for } N_G = 1.$$

The spatial density of the CPCU rain gauge network underlying our daily precipitation estimates varies greatly in time and across CONUS. Therefore, gridded reports of station densities underlying CPCU rain gauge estimates for a representative sample of days between 1 April 2015 to 30 August 2020 are used to estimate basin-specific values of N_G.

Given estimates of σ_{CPCU} and σ_{L4_SM} for each basin, we can numerically estimate the value of SRCS (i.e., $SRCS_{PC}$) you would expect to sample from observed ASM and RC time series if the PC assumption holds and observed SRCS is degraded only by attenuation bias associated with random observation error in either ASM or RC. This is done by first sorting observed pre-storm L4_SM ASM and USGS/CPCU storm-scale RC values (separately) such that their rank-correlation is unity and then adding random independent errors to both daily precipitation values and ASM consistent with the error models described above. By re-ordering the observed time series in the first step, we create ASM and RC time series that are (by construction) consistent with the PC assumption. Therefore, when we subsequently add random independent errors to these re-ordered time series, their rank correlation (i.e., $SRCS_{PC}$) reflects what we would expect from our observed data if the PC assumption was valid. Based on this logic, we can re-sample the Spearman rank correlation between the synthetically perturbed ASM and RC values to obtain a basin-specific estimate of $SRCS_{PC}$.

This approach is repeated 5,000 times to ensure statistical convergence of the generated $SRCS_{PC}$ values. Comparisons between sampled SRCS and $SRCS_{PC}$ values can then be used to evaluate the validity of the PC assumption. If the PC assumption is valid, SRCS and $SRCS_{PC}$

will be approximately equal. However, in cases where any time-varying factor besides ASM significantly impacts RC, SRCS will be less than $SRCS_{PC}$.

Note that our estimates of $SRCS_{PC}$ are conservative in that they likely underestimate the total magnitude of attenuation bias for multiple reasons. First, USGS streamflow observations are assumed to be free of random error. Second, based on the latest SMAP validation results employing longer records (Colliander et al., 2021), the bias-corrected standard error of SSM estimates in the L4_SM product is 0.034 m^3m^{-3} and thus slightly larger than the 0.032 m^3m^{-3} value applied here. Finally, the dual use of the gauge-based CPCU product to estimate RC (from NLDAS-2/CPCU; Section 2.2) and to force the land model in the L4_SM algorithm (Section 2.3) should result in negative error cross-correlation between RC and ASM estimates and thus represents an additional source of SRCS attenuation bias that is missing in our estimates of $SRCS_{PC}$. For example, an overestimation of precipitation in the CPCU product would result in an overestimation of L4_SM ASM and an underestimation of RC values (since RC represents streamflow normalized by precipitation accumulation). The sign contrast in these errors would, in turn, work to spuriously degrade the otherwise positive rank correlation between true ASM and true RC. Therefore, if we could properly account for ASM and RC error cross-correlation in our error modeling, our derived $SRCS_{PC}$ estimates would be even lower. In summary, our error modeling approach is conservative in the sense that our simplifications (described above) likely result in $SRCS_{PC}$ values that underrepresent the actual impact of attenuation bias. The implications of this will be discussed later in Section 5.

4. Land Surface Modeling

The methodology described above in Section 3 will be used to address our first main objective - the observation-based evaluation of the PC assumption. In contrast, land surface modeling, described here in Section 4, is central to addressing our second objective - the assessment of the PC assumption as a viable LSM calibration strategy. Our target LSM for this objective is the Noah with Multi-Parameterization options (Noah-MP) land model (Niu et al., 2011).

4.1 Noah-MP set-up

All Noah-MP simulations are based on Version 7.2 of the NASA Land Information System (LIS) (Kumar et al., 2006) and 15-minute/0.125°-resolution Noah-MP v3.6 integrations between 1 April 2015 and 30 August 2020. Off-line Noah-MP forcing is based on the NLDAS-2 meteorological dataset utilizing North American Regional Reanalysis variables for all fields except precipitation, which instead uses the NLDAS-2/CPCU precipitation dataset described in Section 2.3. Within each study basin, end-of-day (24 UTC) SSM (0-10 cm) and daily (0-24 UTC) runoff totals (i.e., surface runoff plus baseflow) are spatially aggregated to generate a daily, basin-scale time series. Note that the 0-10 cm definition of SSM applied in Noah-MP is slightly deeper than the 0-5 cm depth assumed in the L4_SM product. The impact of this vertical discrepancy will be discussed below.

4.2 Noah-MP configuration ensemble

The Noah-MP LSM is unique in that it contains built-in options to utilize different physical approaches for the representation of land surface water and energy balance processes. Here, we leverage this flexibility to generate a 41-member ensemble of Noah-MP configurations that reflects a range of approaches for representing the link between soil moisture and both runoff and evapotranspiration. Calibration results are based on selecting a single member of this ensemble, separately for each basin, that maximizes a particular calibration objective function (see Section 4.4 below).

Given our emphasis on the representation of runoff, we start with the 16-member Noah-MP configuration ensemble defined by Crow et al. (2019) via the systematic variation of Noah-MP runoff processes. Within this ensemble, separate Noah-MP simulations are generated for all four Noah-MP runoff-physics packages described in Niu et al. (2011): the simplified groundwater (SIM GW) case, the simplified TOPMODEL (SIM TOP) case, the free-drainage (FD) lower-boundary assumption, and the surface runoff parameterization taken from the Biosphere Atmosphere Transfer Scheme (BATS).

For each of these four baseline cases, Crow et al. (2019) selected, and systematically varied, one key parameter to further generate an ensemble of 16 different Noah-MP runoff configurations. For the SIM GW and SIM TOP cases, we selected the TOPMODEL f parameter, which describes the decay of saturated hydraulic conductivity with depth. For the FD case, we selected the *REFKDT* parameter, which modulates the impact of ASM on surface runoff. For the BATS case, we selected the exponential parameter (q), which links the top 2-m soil moisture and surface runoff. In total, four f variations in the SIM GW case, five f variations in the SIM TOP case, four *REFKDT* variations in the FD case, and three q variations in the BATS case were applied to generate the entire 16-member Noah-MP runoff-configuration ensemble. Table S.1 in the Supporting Information and Crow et al. (2019) provide additional details on this ensemble.

Given the overall dominance of evapotranspiration (ET) as a soil water loss mechanism, and the coupling between LSM representations of ET and runoff (Koster and Milly, 1997), the 16-member Noah-MP runoff configuration ensemble of Table S.1 was augmented using an additional 25-member ET-configuration ensemble. The coupling relationship between soil moisture and ET is impacted by a range of LSM parameters and processes. However, as shown in Dong et al. (2020), SSM-ET coupling strength in Noah-MP is highly sensitive to the unitless parameter λ, which controls the nonlinear relationship between soil evaporation stress and SSM. Therefore, a broad range of SSM-ET coupling strengths can be captured by Noah-MP configurations utilizing a corresponding range (i.e., 1, 2, 3, 5, and 10) of λ values (Dong et al., 2020). Therefore, each of these five baseline λ cases was run for the five separate baseline runoff-configuration cases introduced above (i.e., the baseline parameterizations for the SIM GW, FD, and BATS cases, plus two separate FD cases) to generate a 25-member ET-configuration ensemble. For additional details on each configuration within the ET ensemble, see Table S.2 in the Supporting Information and Dong et al. (2020).

This new 25-member ET-configuration ensemble is combined with our earlier 16-member runoff-configuration ensemble to generate a final 41-member Noah-MP ensemble. All calibration results are based on the basin-wise selection of individual Noah-MP configurations within this new 41-member ensemble that maximize one of the objective functions discussed below in Section 4.4. Therefore, this ensemble effectively represents the parameter space for our calibration analysis. Naturally, there exists an extremely wide range of approaches for generating LSM configuration ensembles, and the selection of any single approach is inherently subjective. However, the most important consideration is whether the selected ensemble spans a sufficiently wide range of outcomes to serve as the basis for a robust calibration analysis. In this regard, our 41-member ensemble appears adequate. Across all 617 study basins, the median range of Spearman rank correlation against USGS RC observations within our 41-member ensemble is 0.46. This range falls no lower than 0.13 for any single basin. Therefore, substantial and consistent performance spread is found in all basins between the best and worst Noah-MP configurations contained within our 41-member ensemble.

4.3 Noah-MP spin-up

All members of the final 41-member Noah-MP ensemble are spun-up individually from a cold start on 1 January 2010 until the start of our analysis on 1 April 2015. To demonstrate the adequacy of a five-year spin-up period, we examined Noah-MP configurations utilizing a groundwater model (i.e., SIM GW runoff physics) under the assumption that they possess the most stringent spin-up requirements. Noah-MP runoff and root-zone soil moisture results in these (groundwater model-based) configurations generally stabilized after about five years and were only marginally impacted by further increasing the model spin-up period from 5 to 15 years. This is consistent with Crow et al. (2019) who found that a six-year spin-up period was adequate for examining the relationship between Noah-MP ASM and RC estimates.

4.4 Noah-MP calibration strategies

This section provides a brief description of the LSM calibration strategies applied to the 41-member ensemble of Noah-MP configurations described in Section 4.2 above. For our purposes, the term calibration refers to the selection of a single Noah-MP configuration from this ensemble, separately for each basin, by maximizing one of the three objective functions described below. The result being the definition of a single "calibrated" ensemble member for each individual basin and each individual calibration strategy. Note that the defining characteristic of all three calibration strategies listed below is their exclusive reliance on the continuous L4_SM estimates. As a result, they can all be applied directly (i.e., without extrapolation) to any basin - even highly regulated ones lacking stream gauges.

4.4.1 PC calibration strategy

The key implication of the PC assumption is that basin-scale RC values, available only in gauged/unregulated basins, can be functionally replaced by ASM estimates from the L4_SM product. To examine this possibility, we define the following calibration objective function:

$$F_1 = R_s[\mathbf{ASM}_{\mathrm{L4}}, \mathbf{RC}_{\mathrm{NOAHMP}}]. \tag{4}$$

Here, R_s is a Spearman-rank correlation operator; $\mathbf{ASM}_{\mathrm{L4}}$ is a set of end-of-day antecedent (see Section 2.2) SSM estimates from the L4_SM product; and $\mathbf{RC}_{\mathrm{NOAHMP}}$ is a set of storm-scale RC estimates from a particular Noah-MP configuration (Section 4.2). If the PC assumption holds, selecting a single Noah-MP configuration that maximizes F_1 will simultaneously maximize the correlation of Noah-MP RC estimates versus true values. Hereinafter, the application of (4) as a calibration objective function will be referred to as the "PC calibration" strategy.

4.4.2. SSM calibration strategy

As discussed in Koster et al. (2018), an alternative calibration strategy is maximizing:

$$F_2 = R_p[\mathbf{SSM}_{\mathrm{L4}}, \mathbf{SSM}_{\mathrm{NOAHMP}}] \tag{5}$$

where R_p is the Pearson correlation operator; $\mathbf{SSM}_{\mathrm{L4}}$ are end-of-day SSM (i.e., mean SSM within the final 3-hour window of the UTC day – see Section 2.2) estimates from the L4_SM product; and $\mathrm{SSM}_{\mathrm{NOAHMP}}$ are comparable end-of-day SSM values obtained from a given Noah-MP configuration (Section 4.2). As a result, maximizing F_2 within a particular basin selects Noah-MP configurations that maximize the temporal correlation between L4_SM and Noah-

MP SSM estimates. Note that the complete record of daily SSM estimates is used in (5) - not just the sub-set of days when they describe ASM conditions for a new storm event. Hereinafter, the application of (5) as a calibration objective function will be referred to as the "SSM calibration" strategy. Since it represents an alternative, and more direct, application of L4_SM SSM estimates, the SSM calibration strategy presents a useful baseline for evaluating the PC calibration strategy.

4.4.3 PC+SSM calibration strategy

Combined calibration approaches are also possible including:

$$F_3 = Z(F_1) + Z(F_2) \tag{6}$$

where Z is a normalization function defined as $Z(\mathbf{X}) = [\mathbf{X}\text{-mean}(\mathbf{X})]/\text{std}(\mathbf{X})]$, and $\mathbf{X}$ is a set of (time-invariant) F_1 or F_2 values calculated across all potential Noah-MP configurations for a single basin. Hereinafter, the maximization of (6) will be referred to as the "PC+SSM calibration" strategy.

4.5 Calibration evaluation

The individual Noah-MP configurations selected via the maximization of (4)-(6) are assessed using two different evaluation metrics. The first metric is the Spearman rank correlation (R_s) between the storm-scale RC estimates of a particular Noah-MP configuration and observed RC values; this metric assesses the skill in detecting storm-to-storm variations in runoff efficiency. (Spearman rank correlation is used instead of Pearson correlation to accommodate the potential for modest levels of non-linearity in the relationship between ASM and RC.) Since R_s is blind to the presence of bias in Noah-MP RC estimates, the mean-absolute-error (MAE) of the Noah-MP RC estimates is used as the second evaluation metric.

The resulting R_s and MAE RC evaluation metrics are normalized relative to the best- and worst-performing Noah-MP configurations found in each basin. For each basin i, each calibration function j (i.e., the PC, SSM, or PC+SSM calibration strategies), and each RC evaluation metric (i.e., R_s or MAE), we individually calculate the ratio:

$$F_{C,i,j} = (C_{i,j} - W_i)/(B_i - W_i) \qquad (7)$$

where $C_{i,j}$ is the RC evaluation metric achieved in basin i through the maximization of calibration function j across the Noah-MP configuration ensemble, and B_i and W_i are the best (i.e., lowest MAE or highest R_s) and worst (i.e., highest MAE or lowest R_s) RC evaluation metrics, respectively, across the Noah-MP configuration ensemble for the same basin i.

The endpoints B_i and W_i are calculated via direct comparison to observed RC values, whereas $C_{i,j}$ reflects an evaluation metric obtained without access to streamflow observations - and instead relies wholly on L4_SM ASM estimates. Therefore, $F_{C,i,j}$ = 1 indicates that, even without access to observed RC, a given calibration approach j applied in basin i has accurately identified the single LSM configuration within the Noah-MP ensemble that optimizes a given RC evaluation metric. In the absence of meaningful calibration, all members of the ensemble will be equally likely, and, sampled across all basins i, $F_{C,i,j}$ will have a median value near 0.5 . Therefore, any net upward shift in the spatial distribution of Fc towards one can be regarded as a positive calibration outcome.

5. Results

5.1 Observation-based evaluation of the PC assumption

In this section, we evaluate SRCS and the PC assumption using only observed RC and LS_SM data. Modeling results based on the calibration of the Noah-MP LSM are not discussed until Section 5.2 below. Figure 1 plots observed SRCS values for the medium-scale CONUS basins that meet the selection criteria introduced in Section 2.1. As described above, SRCS is the sampled rank correlation between (pre-storm) ASM values (L4_SM) and storm-scale RC (based on NLDAS-2/CPCU rainfall and USGS streamflow). Observed SRCS values are relatively high with a median value of 0.75. Note that large areas of western CONUS lack coverage due to our inability to sample at least 50 storm events that are not impacted by snow or frozen soil during the historical SMAP period.

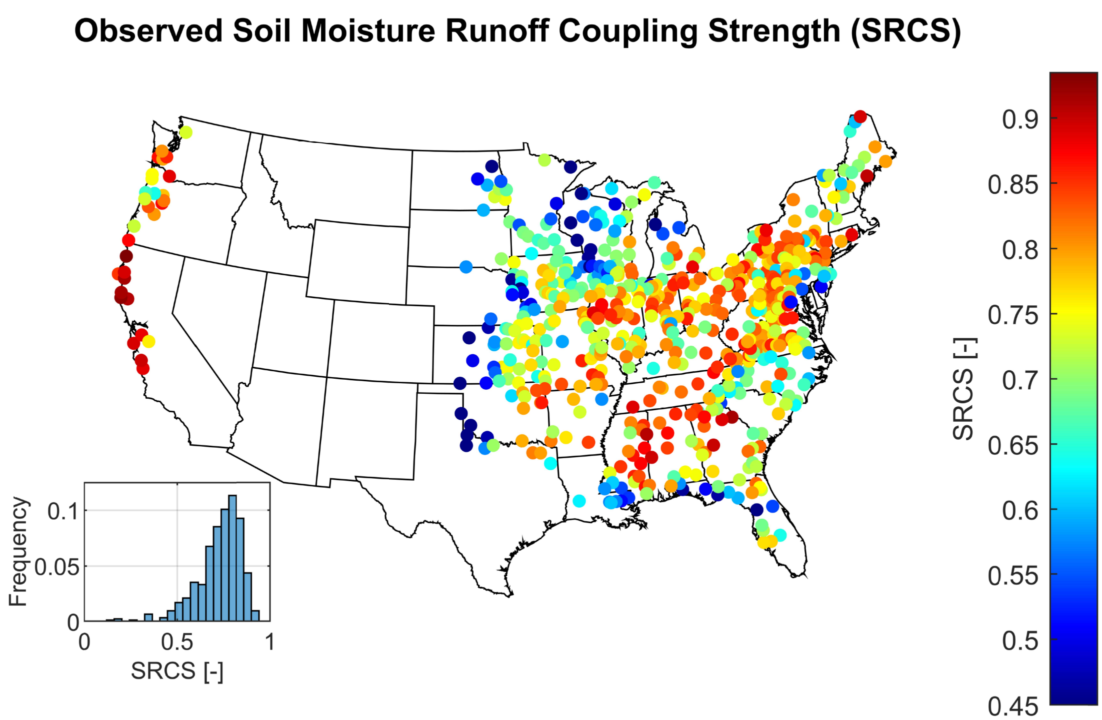


Figure 1. Observed (unitless) SRCS values for all 617 basins meeting the selection criteria described in Section 2.1.

Assuming that random errors in ASM and RC are not positively correlated, sampled SRCS values in Figure 1 will be spuriously reduced by attenuation bias. Therefore, a key question is the degree to which non-unity values of SRCS in Figure 1 reflect the true sensitivity of storm-scale RC variations to non-ASM factors (e.g., the time/space structure of precipitation events) as opposed to the spurious impact of attenuation bias from random errors in the underlying ASM and RC estimates.

As described above, $SRCS_{PC}$ is the hypothetical value of SRCS that we would expect to sample, given attenuation bias from random errors in ASM and RC, if the PC assumption holds and the true rank correlation between ASM and RC is one. As a result, sampled SRCS values that are statistically indistinguishable from $SRCS_{PC}$ imply that the PC assumption is valid. In contrast, SRCS values significantly less than $SRCS_{PC}$ suggest that attenuation bias alone cannot explain the reduction of sampled SRCS below one. This implies that ASM variations cannot explain all observed storm-to-storm variability in RC, and additional factors – presumably related to dynamic land cover variations and/or the exact time/space structure of intra-storm precipitation – must also be considered when modeling the storm-scale RC response. Note that attenuation bias is conceptually distinct from sampling uncertainty in that it does not converge towards zero with increasing sample size.

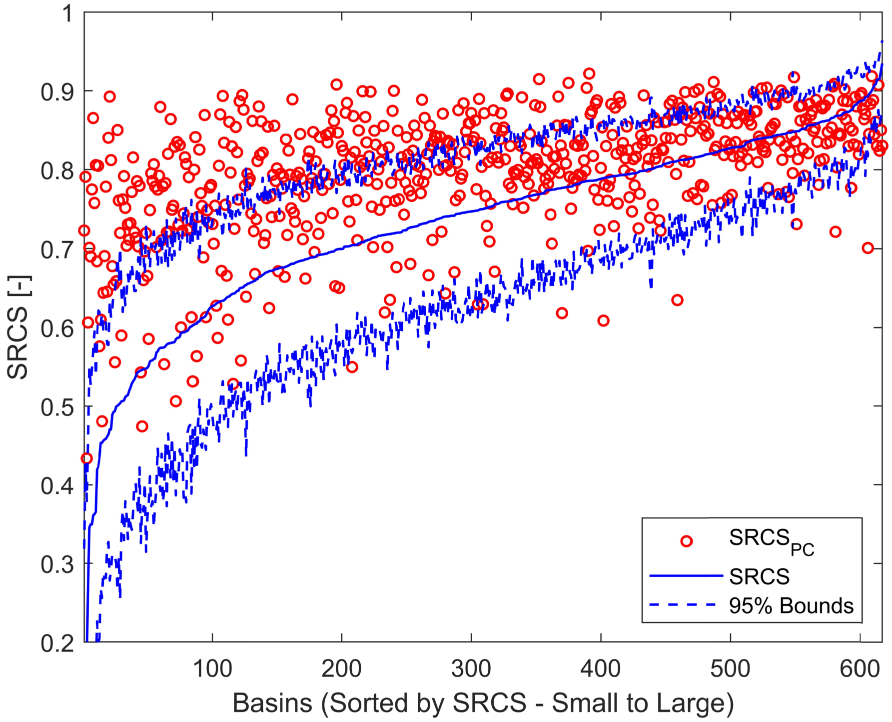


Figure 2. Observed SRCS (solid blue line) and SRCS$_{PC}$ (red circles) for each basin sorted by SRCS. SRCS uncertainty bounds (dashed blue lines) represent the 95% sampling confidence interval calculated based on a 5000-member bootstrapping analysis. ΔSRCS values shown in Figure 3 below are defined as SRCS$_{PC}$ minus SRCS.

Figure 2 compares the observed SRCS with estimated values of SRCS$_{PC}$ for all basins in Figure 1 (sorted by SRCS), along with 95% confidence sampling intervals for SRCS. For 374 of our 617 study basins (~61%), the estimated SRCS$_{PC}$ values fall within these intervals. In these basins, ASM appears to play a dominant role in controlling storm-to-storm variations in RC.

As described in Section 3.2, an analysis of potential bias in SRCS$_{PC}$ estimates suggests that Figure 2 provides a conservative assessment regarding the strength of the PC assumption (i.e., it likely under-represents the fraction of basins where SRCS$_{PC}$ values fall within the 95% uncertainty interval for sampled SRCS values). Furthermore, results in Figure S.1 (see the

Supporting Information) suggest that the impact of inter-day storm events, which could conceivably induce a spurious positive bias into SRCS (see Section 3.2), is negligible.

The difference ΔSRCS = $SRCS_{PC}$ – SRCS quantifies the impact of non-ASM factors on RC temporal variability and, as such, provides a metric for evaluating the validity of the PC assumption. Figure 3a plots ΔSRCS for each basin in Figure 1. As discussed above, ΔSRCS values near zero support the PC assumption while ΔSRCS > 0 implies that factors other than ASM significantly influence RC. For reference, Figure 3b plots AI values for each basin.

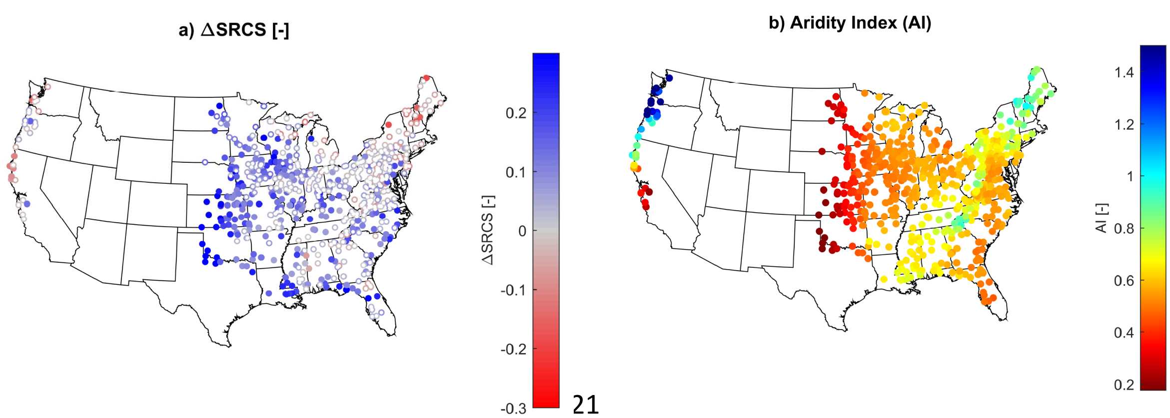


Figure 3. a) ΔSRCS and b) AI for our 617 study basins. Open circles in a) denote basins where ΔSRCS values are not significantly different from zero (at 95% confidence). AI is defined as mean-annual precipitation divided by mean-annual potential evapotranspiration. Therefore, lower AI values indicate generally drier conditions.

ΔSRCS values in Figure 3a are generally statistically insignificant in the relatively humid areas of the northeastern U.S., the mid-Atlantic region, the upper Midwest, and along the spine of the Appalachian Mountains. This is consistent with the expectation that, for humid areas with (at least) modest levels of topographic variability, the dominant runoff mechanism is so-called "saturation-excess" runoff, and the most important factor driving storm-to-storm variations in RC is the fractional area of the basin saturated from below by a dynamic water table (Dunne and Leopold, 1978; Zhao et al., 1980). There is a close conceptual link between this spatial saturation fraction and ASM (Castillo et al., 2003) and, therefore, little observed difference between SRCS and $SRCS_{PC}$.

However, over more arid areas, one would generally expect a lower water table, reduced surface saturation from below, and, therefore, relatively more emphasis placed on non-ASM factors, like the time/space structure of precipitation intensity, when describing storm-to-storm variations in RC (Zhang et al., 2011). This too is reflected in Figure 3a, where generally larger (positive) values of ΔSRCS are noted in more arid (i.e., AI < 0.4) and flatter areas of central CONUS.

Isolated negative ΔSRCS values in northeast and northwest CONUS (see Figure 3a) are difficult to explain physically and likely reflect sampling errors in SRCS estimates, which tend to be larger in northern CONUS basins due to the masking of storm events for snow and frozen soil and/or the possible regional overestimation of L4_SM or CPCU precipitation errors in the statistical models we applied to estimate $SRCS_{PC}$.

Figure 4 summarizes the overall spatial relationship between basin ΔSRCS and AI. For AI > 0.4, ΔSRCS values are often indistinguishable from zero and therefore roughly consistent with the PC assumption. However, for AI < 0.4, ΔSRCS values trend significantly positive, which indicates that the PC assumption is invalid for relatively arid conditions.

Observed outliers in the relationship between ΔSRCS and AI can often be associated with the presence of unusual land surface or geological characteristics in the basin. For example, multiple basins with ΔSRCS values well *above* the general ΔSRCS versus AI trend line in Figure 4 contain very deep glacial soil deposits (either sand or loess) or complex karst sub-surface geology (see symbol labelling in Figure 4). Both characteristics tend to dampen the streamflow response to individual storm events and, therefore, reduce SRCS and increase

ΔSRCS. Conversely, outliers located well below the ΔSRCS versus AI trend line are frequently characterized by a high-degree of anthropogenic stream channelization and/or very-large topographic variability – characteristics that generally enhance the magnitude of streamflow response to storm events. Such enhancement tends to increase SRCS and, as a result, decrease ΔSRCS. Therefore, while AI explains a substantial fraction of observed ΔSRCS spatial variability, land surface and geologic characteristics also play an important role.

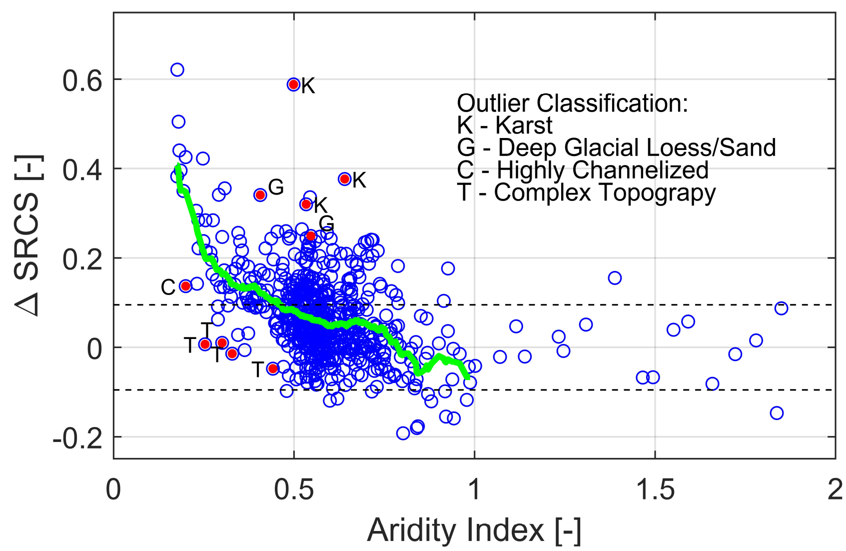


Figure 4. ΔSRCS versus AI values for our 617 study basins. Labelling identifies unusual land surface and geologic conditions in basins that can be visually identified as outliers. The green trend line represents median values within a 0.05-wide moving window applied along the abscissa. Dashed horizontal lines indicate 95% confidence intervals around the null hypothesis that ΔSRCS equals zero. Note that lower values of AI indicate generally drier conditions.

5.2 Noah-MP calibration results

Despite considerable scatter, results in Figures 4 suggest that the PC assumption is potentially applicable for AI > 0.4. In such areas, it may be possible to calibrate LSMs by maximizing the temporal rank correlation between LSM-estimated RC and the L4_SM-based ASM in lieu of RC observations – see (4). Such an approach is particularly attractive in ungauged or heavily regulated basins where meaningful RC observations are not available.

Figure 5 examines this possibility by comparing Noah-MP RC calibration results for the PC, SSM, and PC+SSM calibration strategies introduced Section 4.4. Specifically, the figure shows box-and-whisker plots for the fractional calibration metric F_C defined in (7) across all 617 study basins for both the R_s and MAE RC evaluation metrics.

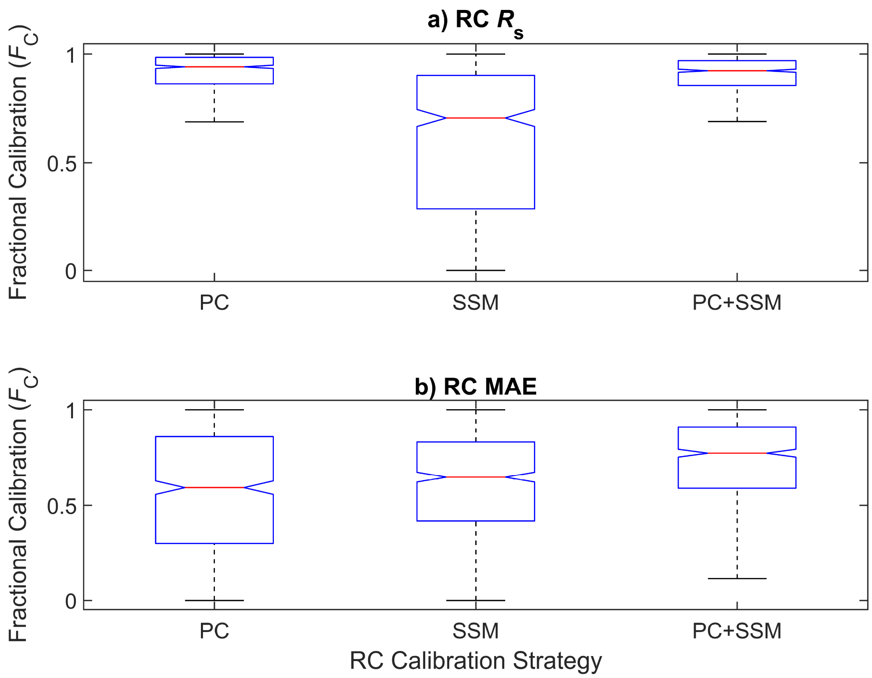


Figure 5. Box-and whisker-plots (summarizing minimum, maximum, median, and inter-quartile spread values) for the basin-to-basin variation of Noah-MP F_C results for all three calibration strategies (i.e., PC, SSM, and PC+SSM) and the a) Spearman correlation (R_s) and b) MAE RC evaluation metrics.

As demonstrated by the concentration of Noah-MP R_s F_C values near one in Figure 5a, the PC calibration strategy consistently identifies a Noah-MP configuration for each basin that has high storm-to-storm Spearman rank correlation with observed RC values. In this regard, the PC calibration strategy is clearly superior to the direct SSM calibration strategy. However, the PC calibration strategy is markedly less successful in identifying Noah-MP configurations that exhibit low RC MAE (Figure 5b). Note that MAE F_C values for the PC strategy in Figure 5b demonstrate only a small net positive shift towards unity. Relative to the PC calibration strategy, the combined PC+SSM calibration strategy provides modestly better MAE results (compare the MAE F_C results for PC and PC+SSM in Figure 5b) while simultaneously preserving its good R_s performance (Figure 5a).

Noah-MP calibration results in Figure 5 are based on utilizing our entire study period (1 April 2015 to 30 August 2020) as both a training and a testing period. Naturally, such overlap raises concerns about overfitting. However, because all three calibration approaches, as well as the B and W values used to derive F_C in (7), are equally impacted by any potential overfitting, Noah-MP calibration results plotted in Figure 5 change very little when re-generated for the case of applying mutually exclusive training and testing periods. For reference, Figures S.2 and S.3 in the Supporting Information provide alternative, but still highly consistent, versions of Figure 5 generated for separate training and testing periods.

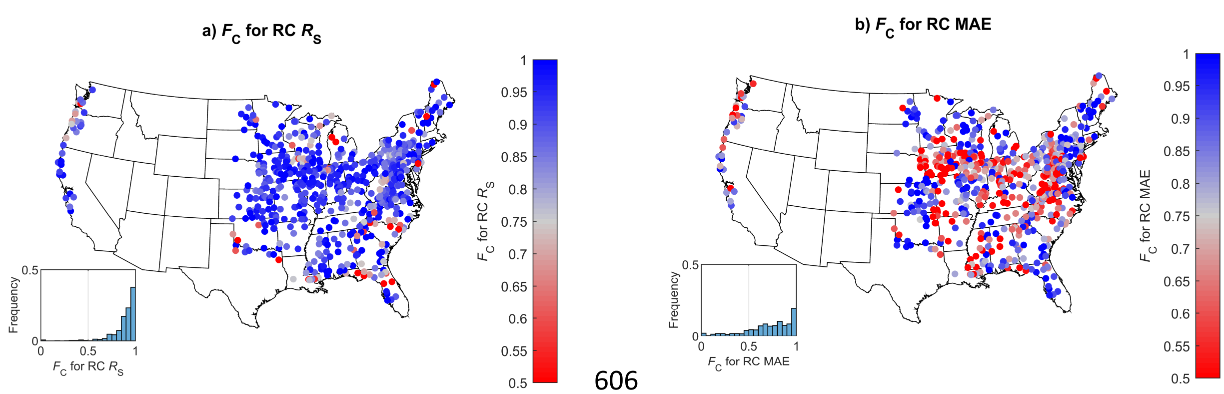


Figure 6. Maps and histograms of Noah-MP F_C results obtained for the PC+SSM calibration strategy based on the a) R_s and b) MAE RC evaluation metrics.

Spatial maps of Noah-MP RC R_s and MAE calibration results (as summarized by F_C) are shown in Figure 6 for the PC+SSM calibration strategy. With a few exceptions, Figure 6a illustrates consistently high Fc results for the Spearman rank evaluation of Noah-MP RC simulations after PC+SSM calibration. Note that some of the, relatively few, poorly performing basins (identified in red in Figure 6a) were previously labeled as outliers in Figure 4 (e.g., highly karst basins in the panhandle region of Florida and basins with extremely deep glacial sand deposits in Michigan). This underscores the challenge of dealing with highly unusual local geology in any calibration strategy not based on local streamflow observations. Conversely, it is surprising that RC R_s results for the PC+SSM strategy are not more clearly degraded in semi-arid regions (e.g., west-central CONUS; Figure 6a), where the PC assumption is known to be tenuous (see Figures 3-4). Given the uncertain physical basis of these results, apparently successful calibration in semi-arid regions should be viewed skeptically.

5.3 Role of RC bias

Despite the overall improvement noted in Figure 5b for the PC+SSM calibration strategy versus the PC strategy, RC MAE values in many basins remain mediocre (grey-shaded) or poor (red-shaded) following PC+SSM Noah-MP calibration (Figure 6b). MAE metric values can be degraded (i.e., increased) by a variety of fit issues including additive bias, multiplicative bias, and poor correlation against true values. As a result, it is worth considering which component of RC MAE is driving these relatively poor results. Additional analysis (see Figure S.4 in the Supporting Information) demonstrates that poor RC MAE F_C results for the PC+SSM calibration strategy are strongly linked to the presence of long-term, basin-specific RC biases in Noah-MP configurations identified as optimal by the PC+SSM calibration strategy.

Figure 7 plots the spatial pattern of long-term RC bias in Noah-MP configurations selected by the PC+SSM calibration strategy. A continuous area of negative RC bias is seen in the northern half of CONUS while a positive RC bias dominates in southern CONUS. The large-scale spatial coherence of this bias pattern suggests that it may be possible to parameterize an additional RC bias-sensitive term using known regional climate or land cover characteristics and add it onto the existing objective function in (6). However, the (correlation-based) PC and PC+SSM calibration strategies considered here are not sensitive to bias and therefore cannot meaningfully contribute to such a term. Instead, alternative, bias-sensitive analysis techniques (e.g., long-term water balance calculations) will be required.

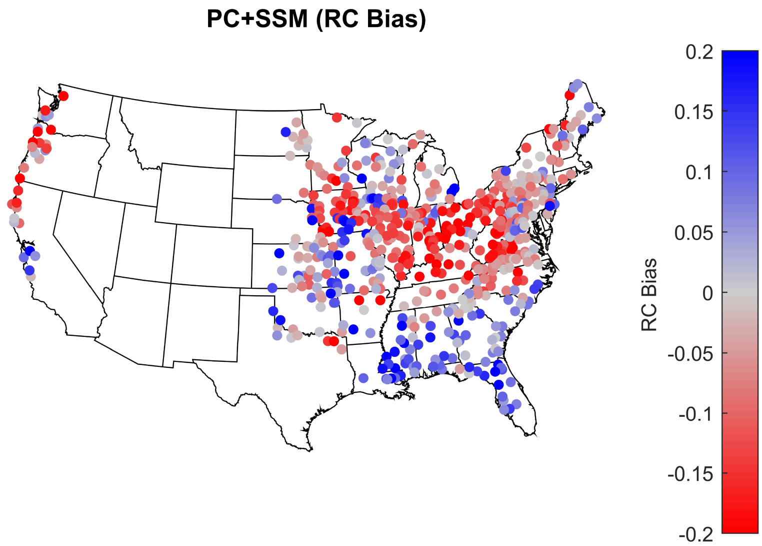


Figure 7. RC bias in Noah-MP configurations identified as optimal by the SSM+PC calibration strategy. Note that the spatial pattern of RC bias drives the mediocre RC MAE results in Figure 6b.

5.4 Sensitivity to methodological changes

5.4.1 Storm-event threshold size

We re-generated Figures 1-3 after raising the storm-event threshold $P_{\min}$ from 10 to 20 mm d^{-1} (not shown). Increasing $P_{\min}$ sharply reduces the number of basins (from 617 to 296) with at least 50 storm events that are not impacted by snow or frozen soil. However, within these 296 basins, which tend to be concentrated in south-central and southeastern CONUS, the median SRCS results for the $P_{\min}$ = 10 mm d^{-1} and 20 mm d^{-1} cases are equal to within two decimal places (0.74). As a result, there is currently no empirical indication that the PC assumption weakens for larger values of $P_{\min}$.

Nevertheless, it is unreasonable to expect the PC assumption to hold indefinitely as $P_{\min}$ is raised without limit. Regardless of underlying ASM levels, storm-scale RC values will eventually converge to unity for extremely intense precipitation events (e.g., $P_{\min}$ approaching or exceeding the expected annual maximum for daily precipitation in a single basin). Note that the relatively short historical record of our analysis places a severe restriction on our ability to sample such extreme events. Further raising $P_{\min}$ to 30 mm d^{-1}, for example, results in only 36 basins with an adequate storm sample size. Therefore, results presented here cannot be extrapolated to characterize basin response to major flooding events.

5.4.2 Vertical support of Noah-MP SSM estimates

As noted above, the vertical support of Noah-MP SSM estimates (0-10 cm) is deeper than the corresponding 5-cm SSM estimates provided by the L4_SM product. The application of a deeper (i.e., > 5 cm) surface soil layer is a common concession in LSMs to numerical difficulties posed by balancing water within excessively thin soil layers. It is possible that this vertical mismatch partially undermines the performance of the SSM calibration strategy posed in (5). However, given the inherent difficulty of matching temporal scales of variability expressed in SSM products obtained from different sources (Shellito et al., 2020), the apparent robustness of the PC and PC+SSM calibration strategies to reasonable variations in surface soil layer depth, and thus temporal SSM memory, is encouraging and likely necessary for their practical application.

5.4.3 Baseflow separation

All calibration results presented above are based on bulk streamflow observations and/or total (i.e., surface + baseflow) runoff estimates provided by Noah-MP. However, owing to the particular importance of predicting fast streamflow response to rainfall, it is common in hydrologic analysis to first remove the baseflow contribution to streamflow. Therefore, a relevant question is whether the, generally good, RC calibration results reflected in Figure 5 for the PC and PC+SSM calibration strategies remain relevant for an application focused only on the fast storm response due to surface runoff.

To remove baseflow from the USGS streamflow observations, we applied the USGS Hydrologic Separation [HYSEP; Sloto and Crouse, 1996] baseflow separation technique using (non-integer) saturation time scales derived from (1) but without the application of the CEIL operator. Note that Noah-MP already separately calculates both surface and baseflow runoff components. Here, unlike above, we based Noah-MP storm-scale runoff estimates on only surface runoff. We then re-generated Figure 5 for this surface-only runoff case, with results shown in Figure 8. A comparison of the two figures reveals no apparent degradation in calibrated RC results associated with considering only surface runoff contributions. In fact, MAE results for the PC+SSM calibration strategy appear to be marginally improved following the removal of baseflow. As a result, the PC or PC+SSM calibration strategies appear to be equally effective at constraining RC estimates based on either surface-only (Figure 8) or total (surface + baseflow; Figure 5) runoff.

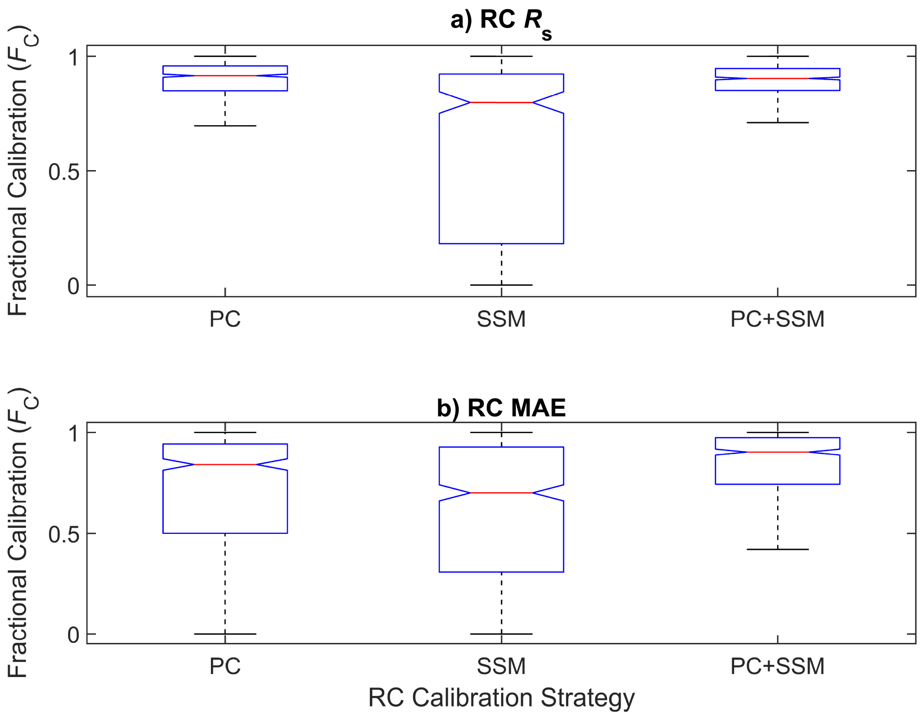


Figure 8. Same as Figure 5 except for only the surface runoff response.

6. Summary and conclusions

Robust external constraints are needed for LSM-based simulation of streamflow in ungauged and/or heavily regulated basins that lack representative streamflow observations. While the potential for remote sensing to provide such constraints is widely acknowledged, the formulation of general physical principles to underlie these constraints remains elusive. Here, we hypothesize that the rank correlation between storm-to-storm variations in ASM and RC is unity (i.e., the PC assumption) and examine if this assumption can serve as a useful calibration principle for obtaining adequate LSM estimates of RC within ungauged basins.

Within a substantial fraction (> 60%) of lightly regulated, median-scale basins in the central and eastern United States, there is indeed no significant evidence that time-varying processes other than ASM significantly impact storm-to-storm variations in RC (Figure 2). This is particularly true for relatively humid (AI > 0.4) basins. To this end, Figures 3a and 4 suggest that, at least over the eastern half of CONUS, the PC assumption is commonly valid.

The apparent magnitude of the rank correlation between ASM and RC suggests utilizing high-quality, spatially continuous ASM estimates provided by the L4_SM product, in lieu of RC observations, as a LSM calibration target in ungauged and/or heavily regulated basins that lack suitable streamflow observations. This approach, summarized in (4) as the “PC calibration strategy”, successfully identifies Noah-MP configurations with relatively good (i.e., high) RC R_s results (Figure 5a). Such a strategy is robust for a range of storm-event rainfall thresholds (i.e., 10-20 mm d^{-1}) and for hydrological applications focused on either total (i.e., surface + baseflow) runoff or surface-only runoff (Figure 8). However, the PC calibration strategy is generally blind to bias in RC estimates – which degrades RC MAE results for its selected Noah-MP configurations (Figures 5b). RC MAE results can be marginally improved by including an SSM-based correlation term in the calibration objective function – via the “PC+SSM” calibration strategy summarized in (6) (Figure 5b). However, the correlation-based calibration strategies we consider here cannot provide a full solution to this bias issue.

There are also geographic limitations to the confident application of the PC assumption. However, these limitations are intuitively associated with known climate and land cover characteristics (Figure 4) – raising hopes that basins where the PC assumption does not hold can be readily identified and masked. We focus on medium-scale (i.e., 200 to 10,000 km^2) basins for the verification of the PC assumption due to the difficulty of defining discrete storm events and screening for streamflow regulation in larger basins. Nevertheless, once verified, we are not aware of any barriers to applying a PC-based calibration strategy to larger basins. Analogously, while the PC assumption can only be verified in lightly regulated basins, there is no reason why in-channel streamflow regulation, in and of itself, should imperil its application to runoff calibration. Finally, given that the integration of SMAP brightness temperature information into the SMAP_L4_SM products has been shown to significantly improve SSM estimation skill in data-poor regions (Dong et al., 2019; Reichle et al., 2021b), we are confident that PC-based calibration strategies can be broadly applied in areas outside of CONUS. However, future work will be needed to confirm this speculation.

Naturally, the evaluation of LSM runoff physics is a multi-objective exercise that requires multiple metrics. While high RC correlation is necessary for adequate LSM correlation, it is certainly not sufficient. For example, additional work is required to define a bias-aware term for (6) that adequately penalizes Noah-MP configurations associated with biased RC estimates (Figure 7). Potential approaches for deriving this term include the application of remotely sensed ET products (obtained, for example, from thermal-infrared remote sensing) or the use of improved river-stage height estimates expected from the upcoming NASA Surface Water Ocean Topography mission (Biancamaria et al., 2016). Likewise, a much longer historical analysis period is required to assess performance for extreme rainfall events likely to cause flooding. Lacking this, care should be taken when applying our proposed calibration approaches to LSMs tasked with flash flood forecasting.

Acknowledgements

This research was supported by grant 80HQTR21T0054 from the NASA SMAP mission ("Using SMAP soil moisture products to improve streamflow forecasting in ungauged basins", PI: W.T. Crow). All USGS, NLDAS-2 forcing data and SMAP soil moisture data used in this study are publicly available and can be accessed online (see in-text data citations). Final soil

moisture and runoff estimates obtained from the ensemble of Noah-MP simulations considered here will be posted in an appropriate on-line data repository following acceptance of this draft manuscript. USDA ARS is an equal opportunity employer.

Work Cited

Biancamaria, S., Lettenmaier, D. & Pavelsky, T. (2016). The SWOT Mission and Its Capabilities for Land Hydrology. *Surveys in Geophysics*. *37*(2), 307-337. https://doi.org/10.1007/s10712-015-9346-y4.

Castillo, V.M., Gómez-Plaza, A. & Martínez-Mena, M. (2003). The role of antecedent soil water content in the runoff response of semiarid catchments: A simulation approach. *J. Hydrol. 284*, 114–130. https://doi.org/10.1016/S0022-1694(03)00264-6.

Chen, M., Shi, W., Xie, P., Silva, V. B. S., Kousky, V. E., Wayne Higgins, R. & Janowiak, J. E. (2008). Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res. 113*, D04110. https://doi.org/10.1029/2007JD009132.

Chen, F., Crow, W.T., Cosh, M.H., Colliander, A., Asanuma, J., Berg, A., Bosch, D.D., Caldwell, T.G., Collins, C.H., Jensen, K.H., Martínez-Fernández, J., McNairn, H., Starks, P.J., Su, Z. & Walker, J.P. (2019). Uncertainty of reference pixel soil moisture averages sampled at SMAP core validation sites. *Journal of Hydrometeorology*. 20, 1553-1569. https://doi.org/10.1175/JHM-D-19-0049.1.

Colliander, A. & coauthors. (2021). Validation of soil moisture data products from the NASA SMAP mission. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 15, 364-392. https://doi.org/10.1109/JSTARS.2021.3124743.

Crow, W.T., Chen, F., Reichle, R.H. & Liu, Q. L. (2017). L-band microwave remote sensing and land data assimilation improve the representation of prestorm soil moisture conditions for hydrologic forecasting. *Geophysical Research Letters. 44*, 5495– 5503 https://doi.org/10.1002/2017GL073642.

Crow, W.T., Chen, F., Reichle, R. Xia, Y. & Liu, Q. (2018). Exploiting soil moisture, precipitation and streamflow observations to evaluate soil moisture/runoff coupling in land surface models. *Geophysical Research Letters*. *45*(10), 4869-4878. https://doi.org/10.1029/2018GL077193.

Crow, W.T., Chen, F., Reichle, R.H. & Xia, Y. (2019). Diagnosing bias in modeled soil moisture/runoff coefficient correlation using the SMAP Level 4 soil moisture product. *Water Resources Research*. *55*, 7010-7026. https://doi.org/10.1029/2019WR025245.

Daly, C., Neilson, R.P. & Phillips, D.L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor. 33*, 140-158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Dong, J. and Crow, W.T. (2018). The added value of assimilating remotely sensed soil moisture for estimating summertime soil moisture - air temperature coupling strength. *Water Resources Research*. *54*, 6072-6084. https://10.1029/2018WR022619.

Dong, J., Crow, W.T., Reichle, R., Liu, Q., Lei, F. & Cosh, M. (2019) A global assessment of added value in the SMAP Level 4 soil moisture product relative to its baseline land surface model. *Geophysical Research Letters*. *46*, 6604-6613. https://doi.org/10.1029/2019GL083398.

Dong, J., Dirmeyer, P. A., Lei, F., Anderson, M.C., Holmes, T.R.H., Hain, C. and Crow, W.T. (2020). Soil evaporation stress determines soil moisture - evapotranspiration coupling strength in land surface modeling. *Geophysical Research Letter.* 47, e2020GL090391. https://doi.org/10.1029/2020GL090391.

Dunne, T. & Leopold, L.B. (1978) *Water in Environmental Planning*. Freeman, New York, 818 p.

Fekete, B.M, Robarts, R.D., Kumagai, M., Nachtnebel, H.-P., Odada, E. & Zhulidov, A.V. (2015). Time for in situ renaissance. *Science. 349*(6249), 685–686. http://doi.org/10.1126/science.aac7358.

Gelaro, R., & coauthors (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version-2 (MERRA-2), *J. Clim.* 30, 5419-5454, https://doi.org/10.1126/science.aac7358.

Hrachowitz, M. & coauthors. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*. *58*(6), 1198-1255. https://doi.rg/10.1080/02626667.2013.803183.

Hutcheon, J.A., Chiolero, A. & Hanley, J.A. (2010). Random measurement error and regression dilution bias. *BMJ. 340*, c2289. https://doi.org/10.1136/bmj.c2289.

Koster, R.D., Liu, Q., Mahanama, S.P.P. & Reichle, R.H. (2018). Improved hydrological simulation using SMAP data: Relative impacts of model calibration and data assimilation. *Journal of Hydrometeorology*. *19*(4), 727-741. https://doi.org/10.1175/JHM-D-17-0228.1.

Koster, R.D. & Milly, P.C.D. (1997). The interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models. *J. Clim. 10*, 1578-1591.

Kumar, S.V., Peters-Lidard, C.D., Tian, Y., Houser, P.R., Geiger, J., Olden, S., Lighty, L., Eastman, J.L, Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E.G. & Sheffield, J. (2006). Land Information System - An interoperable framework for high resolution land surface modeling. *Environmental Modelling & Software*. *21*, 1402-1415. https://doi.org/10.1016/j.envsoft.2005.07.004.

Linsley, R.K., Kohler, M.A., & Paulhus, J.L.H. (1982). *Hydrology for Engineers*. 3rd Edition. McGraw-Hill. New York. 508 p.

Lohmann, D. & coauthors (2004). Streamflow and water balance intercomparison of four land surface models in the North American Land Data Assimilation System (NLDAS). *J. Geophys. Res. 109*, D07S91, https://doi.org/10.1029/2003JD003517.

Lucchesi, R. (2018). File Specification for GEOS FP. GMAO Office Note No. 4 (Version 1.2), 61pp. [Available online at `http://gmao.gsfc.nasa.gov/pubs/`].

Niu, G.-Y. & coauthors (2011), The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, *J. Geophys. Res.* 116, D12109, https://doi.org/10.1029/2010JD015139.

Piepmeier, J. R. et al. (2017), SMAP L-Band Microwave Radiometer: Instrument Design and First Year on Orbit, *IEEE Transactions on Geoscience and Remote Sensing*. *55*(4), 1954–1966, doi:10.1109/tgrs.2016.2631978.

Qiu, J., Dong, J., Crow, W.T., Zhang, X., Reichle, R. & De Lannoy, G. (2021). The added value of brightness temperature assimilation for the SMAP Level-4 surface and root-zone soil moisture analysis over mainland China. *Hydrology and Earth System Sciences*. *25*, 1569-1586. https://doi.org/10.5194/hess-25-1569-2021.

Reichle, R.H. & coauthors (2017), Global assessment of the SMAP Level-4 Surface and Root-Zone Soil Moisture product using assimilation diagnostics. *Journal of Hydrometeorology*. *18*, 3217-3237. https://doi.org/10.1175/JHM-D-17-0130.1.

Reichle, R.H. & coauthors. (2019). Version 4 of the SMAP Level-4 soil moisture algorithm and data product. *Journal of Advances in Modeling Earth Systems*. *11*, 3106-3130. https://doi.org/10.1029/2019MS001729.

Reichle, R., De Lannoy, G., Koster, R.D., Crow, W.T., Kimball, J.S. & Liu, Q. (2020). *SMAP L4 Global 3-hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Geophysical Data, Version 5*. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/9LNYIYOBNBR5. Accessed: 1 December 2020.

Reichle, R.H. & coauthors. (2021a). Soil Moisture Active Passive (SMAP) project assessment report for version 5 of the L4_SM data product, R.D. Koster (Ed). *Technical Report Series on Global Modeling and Data Assimilation*. NASA GMAO, *58*.

Reichle, R. H. & coauthors (2021b). The contributions of gauge-based precipitation and SMAP brightness temperature observations to the skill of the SMAP Level-4 soil moisture product. *Journal of Hydrometeorology*. *22*, 405-424, https://doi.org/10.1175/JHM-D-20-0217.1.

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E.H., Warrach-Sagi, K. & Attinger, S. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrol. Earth Syst. Sci*. *21*, 4323–4346. https://doi.org/10.5194/hess-21-4323-2017.

Schaake, J.C., Duan, Q., Smith, M. & Koren, V. (2000). Criteria to select basins for hydrologic model development and testing. Preprints in 15th Conf. On Hydrology (Long Beach, California, USA, Am. Met. Soc., 10–14 January 2000), Paper P1.8.

Shellito, P.J., Kumar, S.V., Santanello, J.A., Jr., Lawston-Parker, P., Bolten, J.D., Cosh, M.H., Bosch, D.D., Holifield Collins, C.D., Livingston, S., Prueger, J., Seyfried, M. & Starks, P.J. (2020). Assessing the impact of soil layer depth specification on the observability of modeled soil moisture and brightness temperature. *Journal of Hydrometeorology*. *21*(9), 2041-2060. https://doi.org/10.1175/JHM-D-19-0280.1

Sloto, R.A. & Crouse. M.Y. (1996). HYSEP: A computer program for streamflow hydrograph separation and analysis. *U.S. Geological Survey Water-Resources Investigations Report 1996–4040*, 46 p., https://pubs.er.usgs.gov/publication/wri964040.

Song, S. & Wang, W. (2019). Impacts of antecedent soil moisture on the rainfall-runoff transformation process based on high-resolution observations in soil tank experiments. *Water*. *11*(296). https://doi.org/10.3390/w11020296.

United Nations Environment Programme. Middleton, N., & Thomas, D. S. G. (1992). *World atlas of desertification*. London: Edward Arnold.

Villarini, G., Mandapaka, P.V., Krajewski, W.F. & Moore, R.J. (2008). Rainfall and sampling uncertainties: A rain gauge perspective. *J. Geophys. Res. 113*, D11102, https://doi.org/10.1029/2007JD009214.

Xia, Y. & coauthors. (2009). *NCEP/EMC, NLDAS Primary Forcing Data L4 Hourly 0.125 x 0.125 degree V002*, Edited by David Mocko, NASA/GSFC/HSL, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: 1 November 2020, https://doi.org/10.5067/6J5LHHOHZHN4

Xia, Y. & coauthors. (2012). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *J. Geophys. Res.*, *117*, D03110, https://doi.org/10.1029/2011JD016051.

Zhang, Y., Wei, H. & Nearing, M.A. (2011). Effects of antecedent soil moisture on runoff modeling in small semiarid watersheds of southeastern Arizona. *Hydrol. Earth Syst. Sci. 15*, 3171–3179.

Zhao, R.-J., Zuang, Y.-L., Fang, L.-R., Liu, X.-R. & Zhang, Q.-S. (1980). The Xinanjiang model. In *Proceedings of the Oxford Symposium*, Oxford, UK, 15–18 April 1980; pp. 351–356.

Zomer R.J., Bossio, D.A., Trabucco, A., Yuanjie, L., Gupta, D.C. & Singh, V.P. (2007). Trees and water: Smallholder agroforestry on irrigated lands in northern India. *Colombo, Sri Lanka: International Water Management Institute*. pp 45. (IWMI Research Report 122).

Zomer, R.J., Trabucco, A., Bossio, D.A., van Straaten, O. & Verchot, L.V. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agric. Ecosystems and Envir. 126,* 67-80.

Figure 1.

Observed Soil Moisture Runoff Coupling Strength (SRCS)

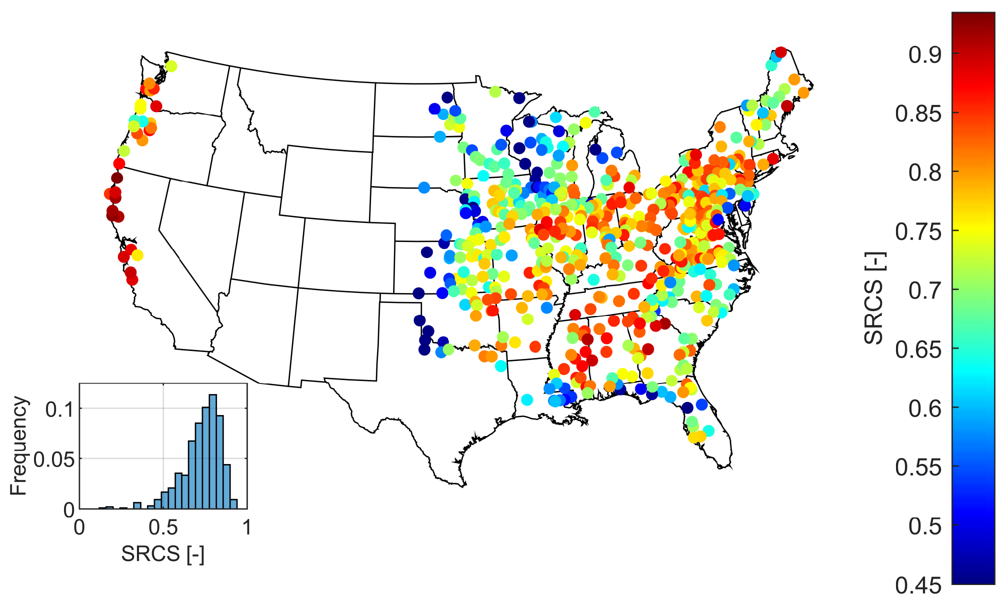
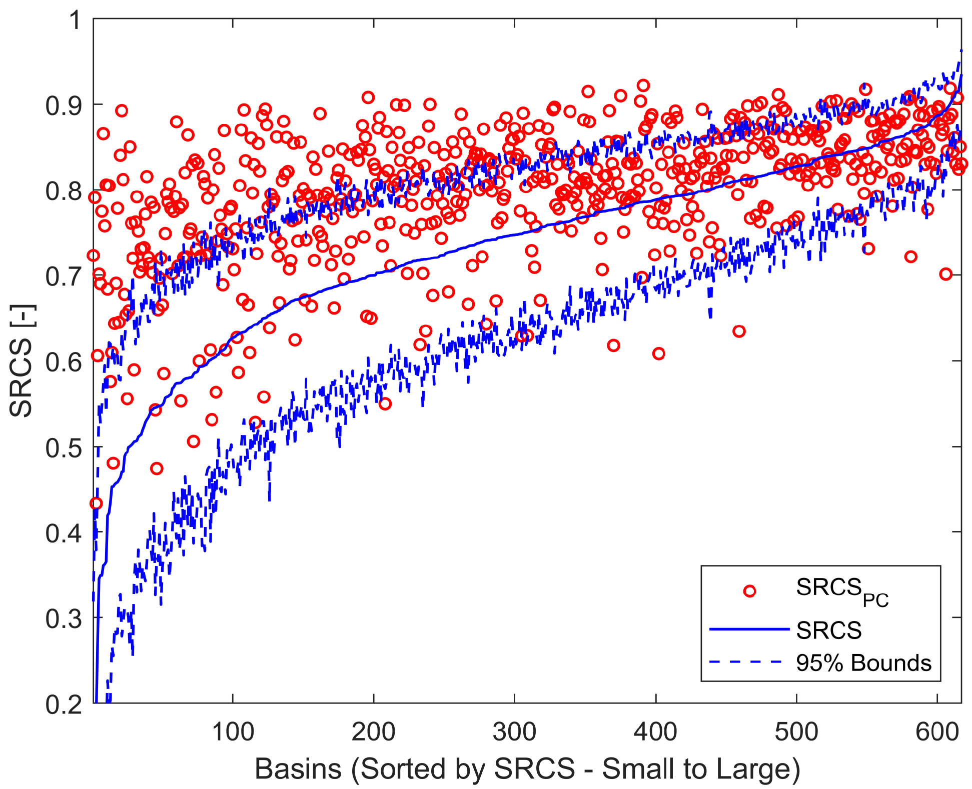


Figure 2.



SRCS [-]
Basins (Sorted by SRCS - Small to Large)
SRCS$_{PC}$
SRCS
95% Bounds
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
100
200
300
400
500
600

Figure 3a.

a) ΔSRCS [-]

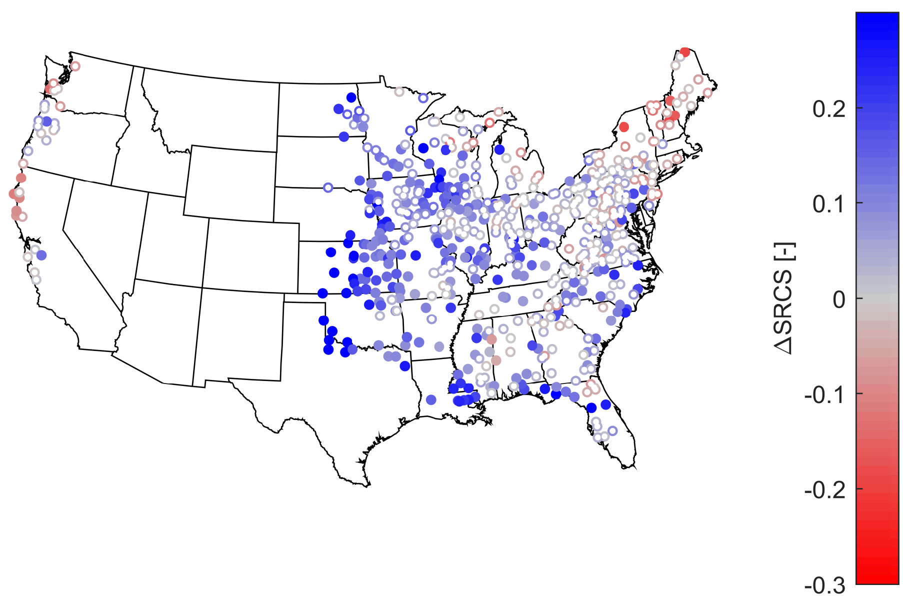


Figure 3b.

b) Aridity Index (AI)

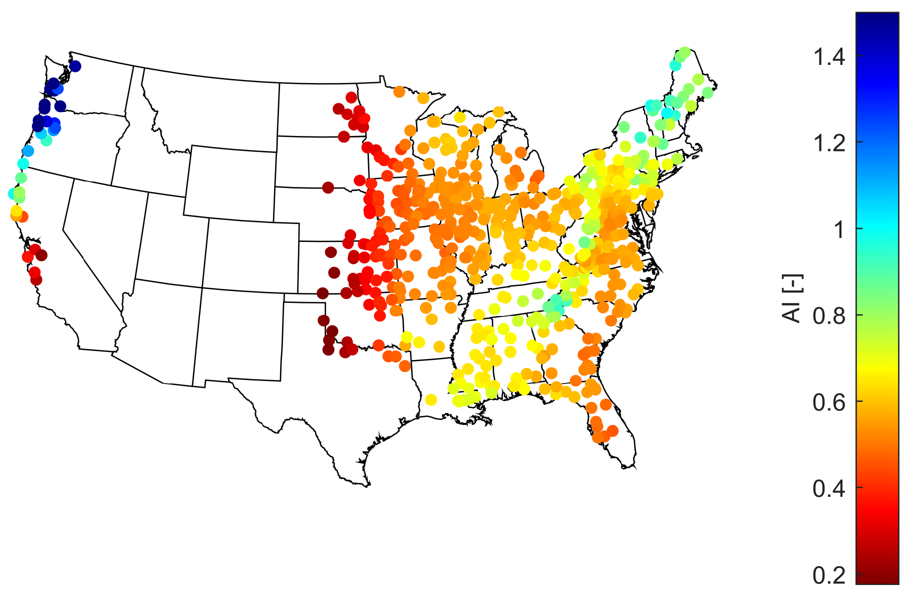
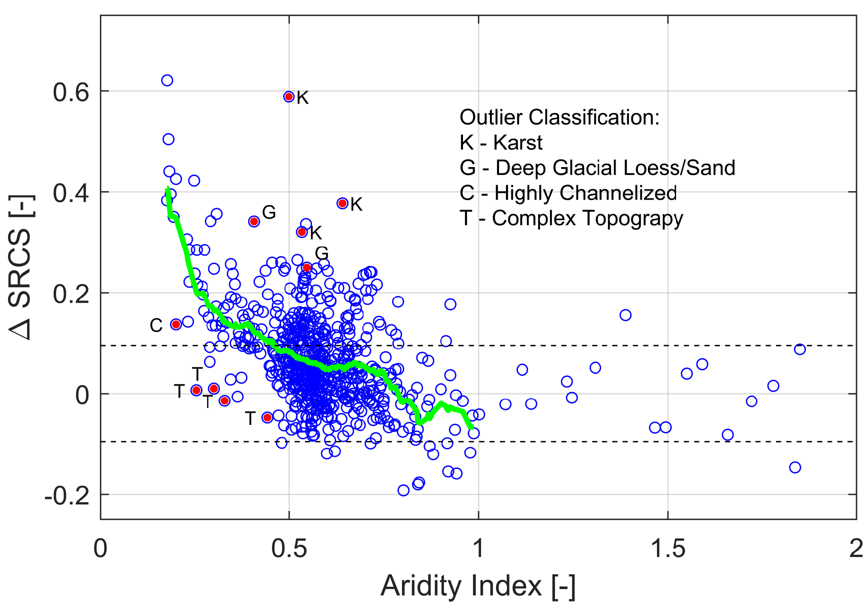
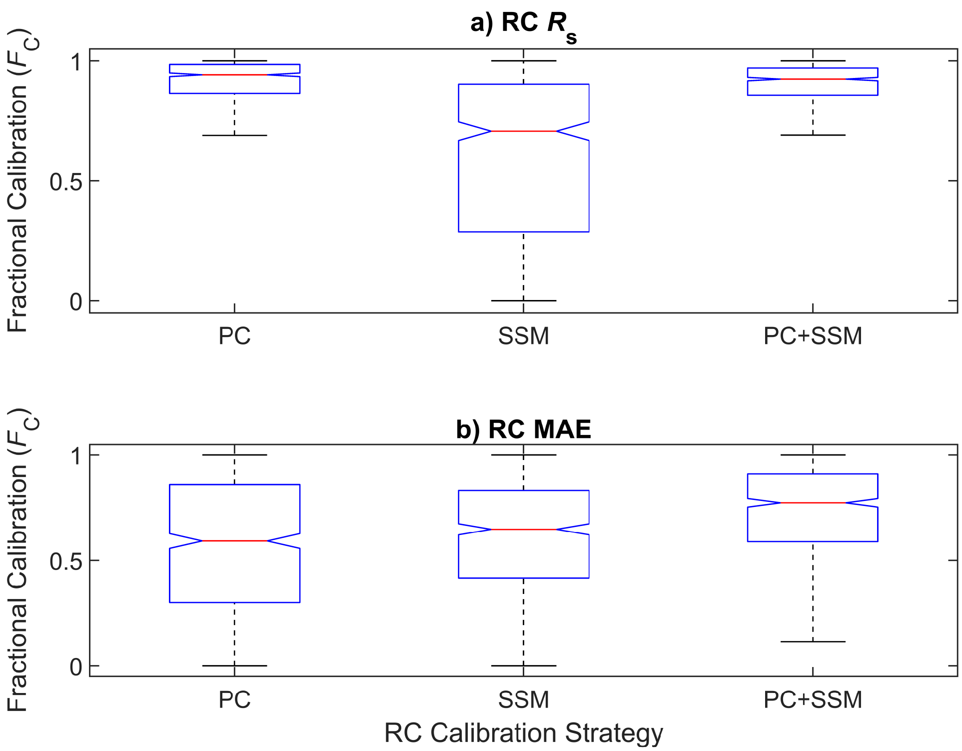


Figure 4.



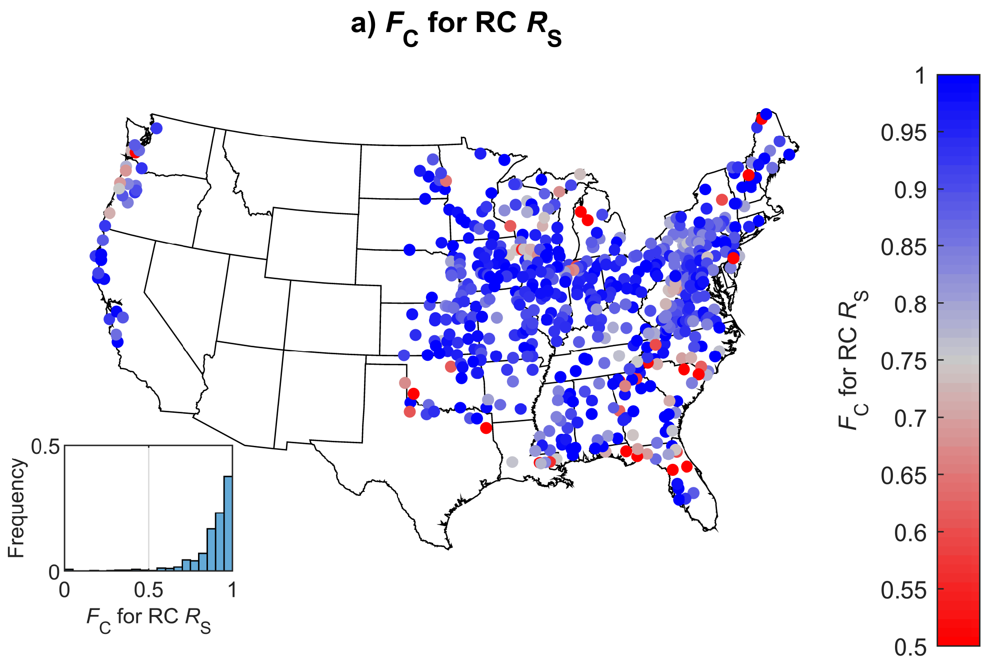
Outlier Classification:
K - Karst
G - Deep Glacial Loess/Sand
C - Highly Channelized
T - Complex Topograpy
Δ SRCS [-]
Aridity Index [-]
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
K
K
K
G
G
C
T
T
T
T

Figure 5.



a) RC R_s
Fractional Calibration (F_C)
1
0.5
0
PC
SSM
PC+SSM
b) RC MAE
Fractional Calibration (F_C)
1
0.5
0
PC
SSM
PC+SSM
RC Calibration Strategy

Figure 6a.



a) F_C for RC R_S
Frequency
0.5
0
0
0.5
1
F_C for RC R_S
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
F_C for RC R_S

Figure 6b.

b) F_C for RC MAE

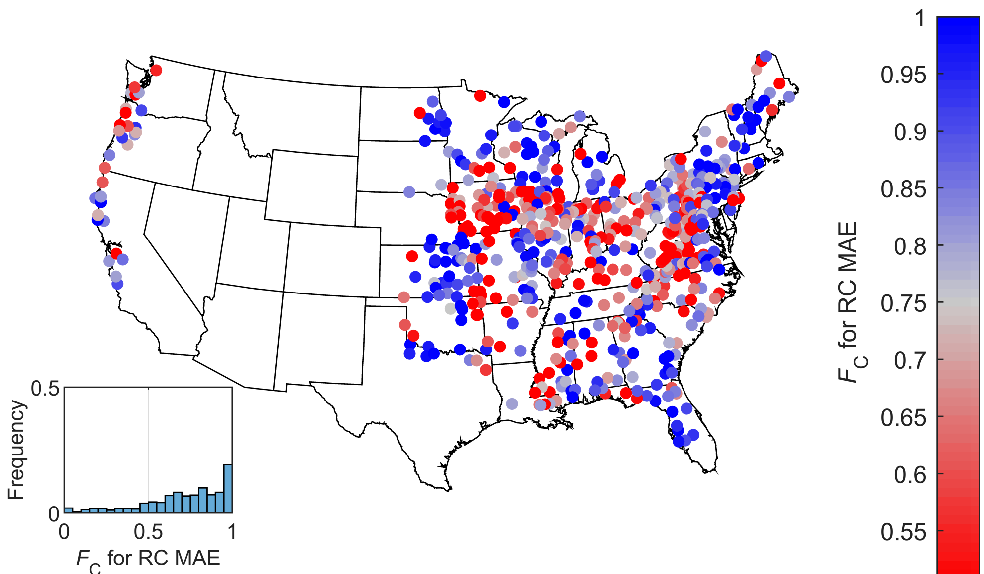


Figure 7.

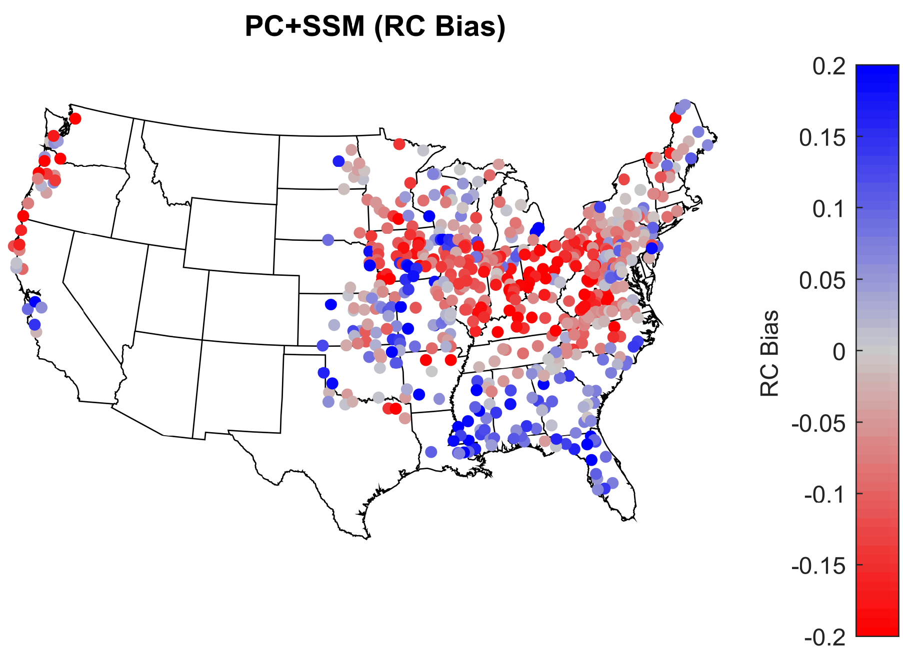
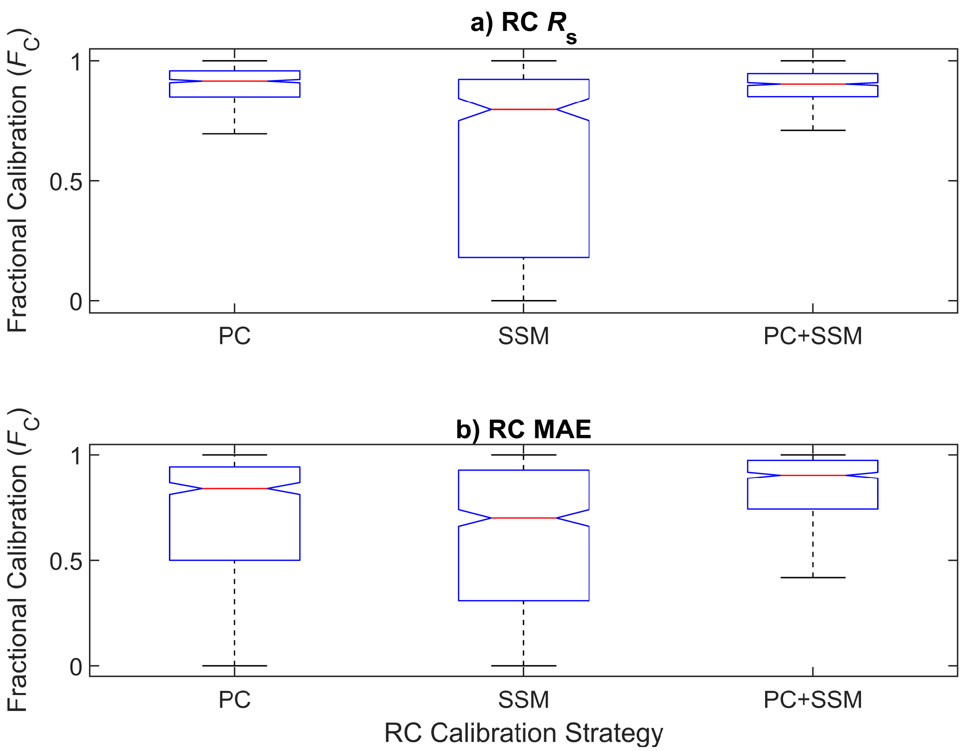

PC+SSM (RC Bias)
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
RC Bias

Figure 8.


a) RC R_s
Fractional Calibration (F_C)
1
0.5
0
PC
SSM
PC+SSM
b) RC MAE
Fractional Calibration (F_C)
1
0.5
0
PC
SSM
PC+SSM
RC Calibration Strategy