Improved Estimates of Pentad Precipitation through the Merging of Independent Precipitation Datasets

Randal D. Koster[1], Qing Liu[1,2], Rolf H. Reichle[1], and George J. Huffman[3]

[1]Global Modeling and Assimilation Office, NASA/GSFC, Greenbelt, MD, USA; [2]Science Systems and Applications, Inc., Lanham, MD, USA; [3]Mesoscale Atmospheric Processes Laboratory, NASA/GSFC, Greenbelt, MD USA

Corresponding author: Randal D. Koster (randal.d.koster@nasa.gov)

Key Points:

- A novel strategy for combining three fully independent precipitation datasets into a single merged precipitation dataset is presented.
- A merged dataset is built from: (i) gauge-based, (ii) satellite-based and (iii) soil moisture retrieval-based precipitation estimates.
- The merged dataset validates better against independent data than does each contributor, benefitting from their complementary strengths.

Abstract

Three independent, quasi-global, gridded datasets of precipitation (a rain gauge-based dataset, the satellite-only component of the NASA Integrated Multi-satellitE Retrievals for Global Precipitation Measurement mission [IMERG] Final Run precipitation product, and precipitation estimates derived from NASA Soil Moisture Active Passive [SMAP] soil moisture retrievals), are objectively combined into a single pentad precipitation dataset at 36-km resolution using a unique approach based on extended triple collocation. The quality of each of the four datasets is then evaluated against independent observations. When a global land surface model at 36-km resolution is integrated four times, once utilizing the merged precipitation forcing and once with each of the three contributing datasets, the near-surface soil moisture variations produced with the merged forcing validate best against independent satellite-based soil moisture fields. In addition, the merged dataset is found to be more consistent, relative to each contributor, with estimates of air temperature variations across the globe. The merged dataset thus appears to draw successfully on the complementary strengths of each contributor: the particularly high quality of the rain gauge-based dataset in areas of high gauge density, the more uniform accuracy across the globe of the IMERG data, and the moderate accuracy, particularly in semi-arid regions, of the soil moisture retrieval-based data.

Plain Language Summary

Obtaining measurements of precipitation across the globe can be challenging. Rain gauges in some ways provide the most accurate measurements, but gauges are absent in many parts of the world, and even where they exist, they only measure precipitation at the gauge itself and therefore may not provide an accurate large-scale average. Satellite-based estimates of precipitation largely overcome these problems, but such data have their own issues, notably a "snapshot" (rather than a time-average) character of the measurements and difficulty associated with interpreting the measured radiances in the presence of complex land surfaces. In the present paper, we use a novel approach to generate a "merged" dataset, one that optimally combines the gauge precipitation information and the satellite-based precipitation information with a third set of estimates derived from soil moisture retrievals. The merged precipitation dataset and each of the three contributors (aggregated here to 5-day averages at a spatial resolution of about 36-km) are then evaluated for consistency with independent geophysical fields. The merged dataset is found to perform best, a clear indication that it takes proper advantage of the complementary strengths of each contributor and, accordingly, that the presented approach for merging the different contributors is indeed viable.

1. Introduction

Precipitation imposes a first-order control on the surface water balance, and accordingly, accurate precipitation forcing is central to accurate hydrological modeling (Larson and Peck, 1974). A hydrological model may comprise well-calibrated and physically sensible treatments of interacting hydrological processes – and may thereby be ready to provide useful estimates of subsurface soil moisture content and transport, groundwater discharge, and large-scale evapotranspiration – but if the precipitation that drives the model is poor, then so too will be its products. The impact of precipitation accuracy on hydrological simulation, which has been explored in numerous studies (e.g., Obled et al., 1994; Renard et al., 2010; Arnaud et al., 2011; Bisselink et al., 2016), has long served as a key motivation for improved precipitation measurement systems (e.g., Sorooshian et al., 2000).

Global hydrological modeling accordingly requires an accurate global dataset of precipitation forcing. Gauge-based global datasets (e.g., Chen et al., 2008; Schneider et al., 2015) have the longest historical legacy and continue to be produced and utilized by the community, and they are generally considered the gold standard where the gauge data are available. The advent of satellite measurements, however, ushered in new strategies for measuring global rainfall variations (Tapiador et al., 2012). In essence, these remote sensing techniques translate radiances from various combinations of hydrometeors, clouds, and water vapor into precipitation rates. The algorithms range from early statistical relationships, such as the Geostationary Operational Environmental Satellite (GOES) Precipitation Index (GPI; Arkin and Meisner, 1987), to the more physically-based Goddard PROFiling algorithm (GPROF; Kummerow et al., 2015; Randel et al., 2020). The former related infrared (IR) cloud-top temperatures to radar rainfall estimates in the Global Atmospheric Research Project (GARP)

Atlantic Tropical Experiment (GATE) in the Atlantic off of West Africa in 1974, while the latter uses a Bayesian scheme to select entries in libraries of vertical profiles of radiative transfer calculations, hydrometeor content, and atmospheric temperature and humidity, each related to a surface precipitation rate, that best match multi-channel passive microwave radiance observations by a sensor.

Note that the gauge-based and satellite-based precipitation datasets have specific advantages and disadvantages. The chief advantage of the gauge-based products is the simple fact that precipitation amounts at the gauge locations are directly measured rather than inferred – the rates obtained at a given station can be considered highly accurate (though still subject to measurement error, e.g., due to undercatch during windy conditions). The usefulness of gauge products at the global scale, however, is limited by: (i) the low density or complete lack of measurement stations in many parts of the world (Kidd et al., 2017), and (ii) the fact that, even in well-gauged regions, the gauges measure precipitation at a point rather than over a large area, so that spatial representativeness errors can significantly degrade the gridded product. In contrast, the data underlying satellite-based precipitation products are far more globally comprehensive, with each data value representing an areal average rather than a point measurement. Furthermore, satellite datasets can potentially offer high spatial and temporal resolution [e.g., 0.1°×0.1°, half-hourly for the Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (GPM) mission (IMERG); see Section 2.1.2]. Satellite-based products, however, are subject to their own disadvantages: (i) since the individual measurements represent snapshots in time, the data are subject to temporal representativeness error, and (ii) the satellites measure radiances that must be converted into precipitation rates using calibrated algorithms, and these algorithms are particularly difficult to apply over heterogeneous land surfaces.

An additional, fully independent approach to deriving global gridded datasets of precipitation forcing has recently been garnering attention. The SM2RAIN algorithm (Brocca et al., 2013, 2014) interprets time variations in remotely-sensed soil moisture retrievals in terms of the precipitation rates that forced them (see Section 2.1.3). As with the other satellite-based products, the soil moisture retrievals underlying SM2RAIN represent areal averages with extensive global coverage. However, because soil moisture integrates, in a sense, the impacts of precipitation over time, the "snapshot" issue limiting the other satellite-based products is less problematic. The SM2RAIN algorithm, of course, has its own important limitations, including: (i) an inability to capture high intensity precipitation estimates, for which liquid precipitation might run off directly rather than infiltrate the soil, (ii) coarse time resolution (as controlled by the revisit time of the satellite) and spatial resolution, (iii) errors associated with the unknown timing of the precipitation between the soil moisture retrievals, and (iv) poor or no estimates in areas with snow, frozen ground, or dense vegetation.

We focus in this paper on the benefits of combining these three distinct and fully independent global precipitation dataset types – datasets based on rain gauge measurements, satellite measurements of cloud and water vapor properties, and satellite measurements of soil moisture – into a single merged dataset that capitalizes on the relevant advantages of each. Because the merging process combines the three contributors optimally based on estimates of their relative accuracies, the merged dataset should, in principle, prove superior to each contributor on its own.

We use a triple collocation-based approach (see Section 2.2) to merge the three contributing precipitation datasets. The approach is simpler than some existing approaches (e.g., Beck et al. 2017) but has the advantage of offering a uniquely intuitive estimation of the relative accuracy of each dataset. We thus view our study as complementing existing work. Triple collocation, which has been used extensively in the geosciences (e.g., Stoffelen, 1998), is made viable here by the independence of the errors in the three contributing precipitation datasets. Triple collocation has in fact already been used for merging precipitation datasets; Dong et al. (2020) used it to merge satellite-based, reanalysis, and SM2RAIN precipitation data into a single merged precipitation product that they then evaluated against a gauge-based precipitation dataset in Europe. These authors found that their merged product indeed validates better against the gauge-based dataset than does any of their contributing datasets individually, illustrating clearly the potential effectiveness of the approach.

Here we employ the same general strategy as Dong et al. (2020), but with two important differences. First, instead of using reanalysis precipitation as one of the three contributors, we use a gauge-corrected weather analysis dataset (in most areas, see Section 2.1.1); because our goal is to produce the most accurate precipitation dataset possible for global hydrological modeling, we want the merged product to take full advantage of the gauge information where it exists – we want to ingest the gauge data into our merged product to the fullest extent possible. Second, we extract the weights used for the merging in a way that is, to our knowledge, unique (section 2.2). Our specific approach may have applicability to data merging exercises in general.

The various precipitation datasets we examine differ in their spatial and temporal resolutions. To make the interpretation of this first study more straightforward, we focus here on the information content of each dataset at a relatively coarse spatial (~36-km x 36-km) and temporal (5-day average) resolution. That is, after coarsening each contributing dataset as necessary to these resolutions and combining them into a single merged dataset, we determine whether the merged dataset validates better against independent data than do any of the coarsened contributors. (The added benefit obtained from the higher resolution information available with some datasets will be addressed in a future study.) Because the soil moisture retrievals used are unavailable during snow periods, we also limit our analysis here to the boreal warm season of May–September. For our main validation exercise, we evaluate the increase in accuracy achieved (relative to an available, independent global dataset of soil moisture retrievals) when the merged data rather than the individual contributors are used to generate soil moistures in a hydrological modeling system.

Finally, note that we will not evaluate absolute magnitudes of precipitation. Such an evaluation is intractable given that we are already using the best precipitation data available to produce the merged dataset, and any fully independent large-scale "truth" we did come across would be subject to the limitations noted above. While we might instead attempt to get at precipitation magnitudes by comparing the streamflow totals generated with our hydrological modeling system against observed streamflow totals, this would almost certainly reflect more on the accuracy of the land model than on the precipitation inputs themselves. In essence, in this study, the long-term (climatological) averages of the magnitudes of our merged precipitation data will be forced to agree with those of the gauge data. Our focus for evaluation will instead be on the time variability of the estimated precipitation amounts. The precise timing and relative magnitudes of events in a precipitation time series are indeed key to the overall characterization of hydrological variability and to the modeling of interactions between the land surface and the rest of the climate system. The time variability of precipitation, our focus here, will draw from all three contributors and should be most accurate within the merged dataset.

Our overall approach – the contributing datasets, the merging procedure, and the validation methodology – is described in Section 2. Our inferences regarding the relative accuracy of the different contributing datasets and our evaluation of the merged precipitation product against available global data are provided in Section 3. Section 4 provides further discussion, and Section 5 provides an overall summary.

2. Data and Methods

2.1 Precipitation datasets.

2.1.1 *Gauge/Analysis.*

The first precipitation dataset considered, the "Gauge/Analysis" dataset, consists of the precipitation data used in the production of the SMAP Level 4 soil moisture product (Reichle et al., 2017b). For the most part, it can be considered a rain gauge dataset; over most of the globe, the Gauge/Analysis data are derived from the 0.5-degree daily Climate Prediction Center Unified (CPCU) rain gauge precipitation product (Xie et al., 2007; Chen et al., 2008), with consideration of the different gauge reporting times (Reichle and Liu, 2014; Reichle et al., 2021). The exceptions are Africa and the high latitudes, where the CPCU gauge coverage is considered too poor for SMAP Level 4 production. In Africa and north of 62.5°N, the Gauge/Analysis data consist of precipitation data produced by the National Aeronautics and Space Administration (NASA) Global Earth Observing System (GEOS) Forward Processing (FP) weather analysis and forecasting model (Lucchesi, 2018). These latter data thus consist of model-generated precipitation amounts from a full atmospheric model "analysis" constrained heavily by assimilated observations of atmospheric temperature, humidity, winds, etc. Between 42.5°N and 62.5°N, a tapered blend of the gauge and analysis data is used. Note that as part of the construction of the Gauge/Analysis dataset, both the CPCU rain gauge data and the GEOS FP data were scaled so that their monthly climatologies matched those of Version 2.2 of the Global Precipitation Climatology Project (GPCP; Adler et al., 2003).

Reichle et al. (2017a, 2019) provide a comprehensive description of this composite precipitation dataset. Prior to merging it with the other two precipitation datasets, we regrid the data to the ~36-km x ~36-km SMAP EASE grid (Brodzik, 2012) and temporally average them to 5-day means. The spatial regridding is performed through bilinear interpolation.

2.1.2 *IMERG.*

IMERG is a U.S. GPM Science Team precipitation product. IMERG provides half-hour, 0.1° x 0.1° global gridded data in three "Runs"—Early (4h after observation time), Late (14h after observation time), and Final (3.5 months after observation time). The algorithm intercalibrates, merges, and interpolates satellite microwave precipitation estimates as well as IR satellite estimates (intercalibrated to the microwave estimates) and precipitation gauge analyses at fine time and space scales for the period June 2000 to present over the globe (Huffman et al., 2020). In this study we use the satellite-only (precipitationUncal) data field in the Version 6 Final Run (which is effectively the Late Run product), meaning that the estimates do not include explicit monthly gauge information and are thus independent of the gauges underlying the Gauge/Analysis data in this study. The IMERG precipitation data are aggregated to the ~36-km x ~36-km SMAP EASE grid and temporally averaged to 5-day means.

2.1.3

SM2RAIN-based Rainfall Estimates. Soil moisture tends to increase during a precipitation event, and the size of the increase is closely related to the precipitation volume. An analysis of satellite-based soil moisture retrievals at a given site should therefore contain information on the precipitation falling at that site. This idea was developed and explored extensively by Brocca et al. (2013, 2014), who successfully derived rainfall estimates from soil moisture retrievals produced by the Advanced Scatterometer (ASCAT) and other sensors using their SM2RAIN algorithm.

Soil moisture retrievals based on L-band brightness temperature measurements are now available from the European Space Agency's SMOS mission (Kerr et al., 2010) and NASA's Soil Moisture Active Passive (SMAP) mission (Entekhabi et al., 2010). These L-band soil moisture retrievals represent conditions in the top 5 cm of soil, a depth ~4-5 times greater on average than that represented by, e.g., ASCAT measurements. The idea that this greater depth is intrinsically more appropriate for SM2RAIN-based precipitation estimation was tested by Koster et al. (2016), who quantified rainfall time series across the globe from SMAP, SMOS, and ASCAT retrievals and then compared them to an established rain gauge-based precipitation product. They found that the L-band retrievals did indeed perform significantly better and that the SMAP data provided the best rainfall estimates.

The SM2RAIN precipitation estimation approach continues to be developed and applied in numerous parts of the globe (e.g., Tarpanelli et al., 2017; Ciabatta et al., 2018; Chiaravalloti et al., 2018). The version of the algorithm utilized here is that described by Koster et al. (2018), which features an empirically-fitted soil moisture loss function that varies in space (Koster et al., 2017). We use SMAP soil moisture retrievals from May-September in 2019-2020 (Release ID R17000) to fit the loss functions (working around a ~5-week SMAP data gap during June-July of 2019), and we then apply the SM2RAIN algorithm to the SMAP retrievals to compute the SM2RAIN-based daily precipitation time series covering May-September of 2015-2018. Note that these time series, while daily, have identical values between the satellite's 3-4 day revisit time. Finally, we aggregate the daily values to pentad averages.

2.2 Merging Approach: Extended Triple Collocation

If the time series of a given variable is estimated in three different ways, using three independent sets of measurements (in particular, measurements with independent errors), the standard deviation of the errors associated with each of the three estimates can be quantified using triple collocation (Stoffelen, 1998). Extended triple collocation (McColl et al., 2014) adds to the theory, providing the means to estimate the correlation $\rho_{X,Truth}$ between each time series X and the unknown "truth" time series – another useful accuracy metric. As discussed below, we use relative values of $\rho_{X,Truth}$ to determine the weights needed to combine the three precipitation datasets into a single merged dataset, with higher weights naturally assigned to the datasets deemed, through $\rho_{X,Truth}$, to be more accurate. These weights, of course, vary with location.

Triple collocation, however, implicitly assumes a Gaussian distribution of the errors in the time series considered, and precipitation is far from Gaussian; even when the precipitation is averaged into pentads, the distributions tend to have a large and positive skew, the pentad precipitation has a nonzero probability of being exactly zero, and precipitation can never be negative. Given these violations of the triple collocation assumptions, the application of triple collocation to raw precipitation time series is arguably difficult to justify. To address the skewness issue, we take the natural logarithm of this average pentad precipitation after first applying a small minimum threshold, set to 1% of the local mean warm-season (May-September) pentad precipitation rate. (The threshold is applied to avoid obvious problems with computing logarithms of zero. Any precipitation rate falling below this threshold is reset to the threshold prior to computing its logarithm.) Lognormal distributions are known to characterize precipitation better than do normal distributions (e.g., Kedem and Chiu, 1987), a fact confirmed by numerous spot-checks with our own pentad data. There still remains, however, the violation of the Gaussian assumption associated with zero precipitation. In addition, the required independence between the datasets may be violated (probably only slightly) in Africa and high latitudes, given that the Analysis precipitation used in these particular regions may be affected by some of the same satellite radiance observations used to derive the IMERG retrievals. Still another possible violation of the triple collocation framework involves the potential presence of significant seasonal cycles in the time series and in the errors (Draper et al., 2013), even for the limited (May-September) timeframe considered here. Given various technical issues (e.g., associated with our use of logarithms – we cannot take a logarithm of a negative anomaly, and the alternative approach of computing anomalies of a time series of logarithms has little physical meaning), we are applying the triple collocation analysis to the original logarithm time series rather than to scaled anomalies relative to a seasonal climatology and are thus disregarding the fact that seasonal cycles could imprint themselves inappropriately on the results.

In other words, even with the use of logarithms, we still face several potential violations to the triple collocation framework. We will nevertheless show in Section 3 that our use of triple collocation leads to a viable merging of the contributing datasets.

At a given grid cell, prior to taking logarithms, we first scale each precipitation time series to have the same long-term May–September mean (that of the Gauge/Analysis data), our goal being to focus on (and find the optimal merging of) the time-variation information contained within each dataset (see Section 1). Letting $L_G(t)$, $L_I(t)$, and $L_S(t)$ represent the time series of the logarithms of the scaled pentad precipitation rates from the Gauge/Analysis, IMERG, and SM2RAIN-based precipitation datasets, respectively, we compute the correlations between each possible pairing:

$$\rho_{GI} = \mathrm{Corr}(\ L_G(t), L_I(t)\ ) \tag{1}$$

$$\rho_{GS} = \mathrm{Corr}(\ L_G(t), L_S(t)\ ) \tag{2}$$

$$\rho_{IS} = \mathrm{Corr}(\ L_I(t), L_S(t)\ ). \tag{3}$$

Now let $\rho_{G,Truth}$, $\rho_{I,Truth}$, and $\rho_{S,Truth}$ represent the temporal correlations between the unknown truth and, respectively, the logarithms of the Gauge/Analysis, IMERG, and SM2RAIN-based pentad data. We use the triple collocation framework to estimate:

$$\rho_{G,Truth} = \{\ \rho_{GI}\ \rho_{GS}\ /\ \rho_{IS}\ \}^{1/2} \tag{4}$$

$$\rho_{I,Truth} = \{\ \rho_{GI}\ \rho_{IS}\ /\ \rho_{GS}\ \}^{1/2} \tag{5}$$

$$\rho_{S,Truth} = \{\ \rho_{GS}\ \rho_{IS}\ /\ \rho_{GI}\ \}^{1/2} \tag{6}$$

See McColl et al. (2014) for further information; these equations are essentially a simplified version of their equation (9). Note that due to sampling error, ρ_{GS}, ρ_{IS}, or ρ_{GI} could be small and negative; we enforce a minimum value of 0.01 for each prior to using them in (4)-(6).

We convert these $\rho_{G,Truth}$, $\rho_{I,Truth}$, and $\rho_{S,Truth}$ values into the weights used to generate the merged product by utilizing a unique (to our knowledge) approach. Consider the general case of three measurement time series $X_1(t)$, $X_2(t)$, and $X_3(t)$ for a given variable that have independent errors and that are normally distributed with zero mean and unit variance. Representing the true (and unknown) standardized time series for the variable as Truth(t), also assumed to be normally distributed with zero mean and unit variance, we can write (e.g., Bras and Rodriguez-Iturbe, 1985):

$$X_1(t) = \rho_{1,\mathrm{truth}}\ \mathrm{Truth}(t) + (1 - \rho_{1,\mathrm{Truth}}{}^2)^{½}\ \varepsilon_1(t) \qquad (7)$$

$$X_2(t) = \rho_{2,\mathrm{Truth}}\ \mathrm{Truth}(t) + (1 - \rho_{2,\mathrm{Truth}}{}^2)^{½}\ \varepsilon_2(t) \qquad (8)$$

$$X_3(t) = \rho_{3,\mathrm{Truth}}\ \mathrm{Truth}(t) + (1 - \rho_{3,\mathrm{Truth}}{}^2)^{½}\ \varepsilon_3(t), \qquad (9)$$

where $\rho_{1,\mathrm{Truth}}$, $\rho_{2,\mathrm{Truth}}$, and $\rho_{3,\mathrm{Truth}}$ are the correlations between the three measurement time series and the unknown truth, and $\varepsilon_1(t)$, $\varepsilon_2(t)$, and $\varepsilon_3(t)$ are independent and normally distributed random variables with zero mean and unit variance. Our goal is to find the weights W_1, W_2, and W_3 that, when used to compute a merged time series, Merged(t):

$$\mathrm{Merged}(t) = W_1\ X_1(t) + W_2\ X_2(t) + W_3\ X_3(t), \qquad (10)$$

maximize the value of the temporal correlation $\rho_{\mathrm{Merged,Truth}}$ of the merged data with the truth. Each of the weights found would be a function of the correlations $\rho_{1,\mathrm{Truth}}$, $\rho_{2,\mathrm{Truth}}$, and $\rho_{3,\mathrm{Truth}}$ already established through triple collocation. Note that this is distinct from the approach of Dong et al. (2020), who focused on minimizing errors rather than maximizing the correlation against the unknown truth.

Finding analytical expressions for W_1, W_2, and W_3 by searching for maxima of the analytical representation of Corr[Merged(t),Truth(t)] quickly becomes intractable. A brute-force Monte Carlo approach, however, is well-suited to the problem. Using a random number generator, we generate a time series Truth(t) with zero mean and unit variance, and, using (7)-(9) along with prescribed values of $\rho_{1,\mathrm{Truth}}$, $\rho_{2,\mathrm{Truth}}$, and $\rho_{3,\mathrm{Truth}}$, we construct artificial time series of $X_1(t)$, $X_2(t)$, and $X_3(t)$ that are fully consistent with this artificial truth. We then test all combinations [W_1, W_2, W_3] and find the particular combination that maximizes Corr[Merged(t),Truth(t)]. This exercise determines the values of W_1, W_2, and W_3 for the particular combination of $\rho_{1,Truth}$, $\rho_{2,Truth}$, and $\rho_{3,Truth}$ examined. By design, the correlation between the merged dataset constructed with these values and the unknown truth equals or exceeds $\rho_{1,Truth}$, $\rho_{2,Truth}$, and $\rho_{3,Truth}$ in this idealized analysis.

We illustrate the weights so generated in Figure 1. Weights are computed for all combinations of $\rho_{1,Truth}$, $\rho_{2,Truth}$, and $\rho_{3,Truth}$ values in increments of 0.01; Figure 1 only shows the sensitivity of the weights to $\rho_{1,Truth}$ and $\rho_{2,Truth}$ for a few selected values of $\rho_{3,Truth}$. For presentation purposes, a 5-point boxcar smoother was applied to the fields to remove a small amount of sampling-related noise.

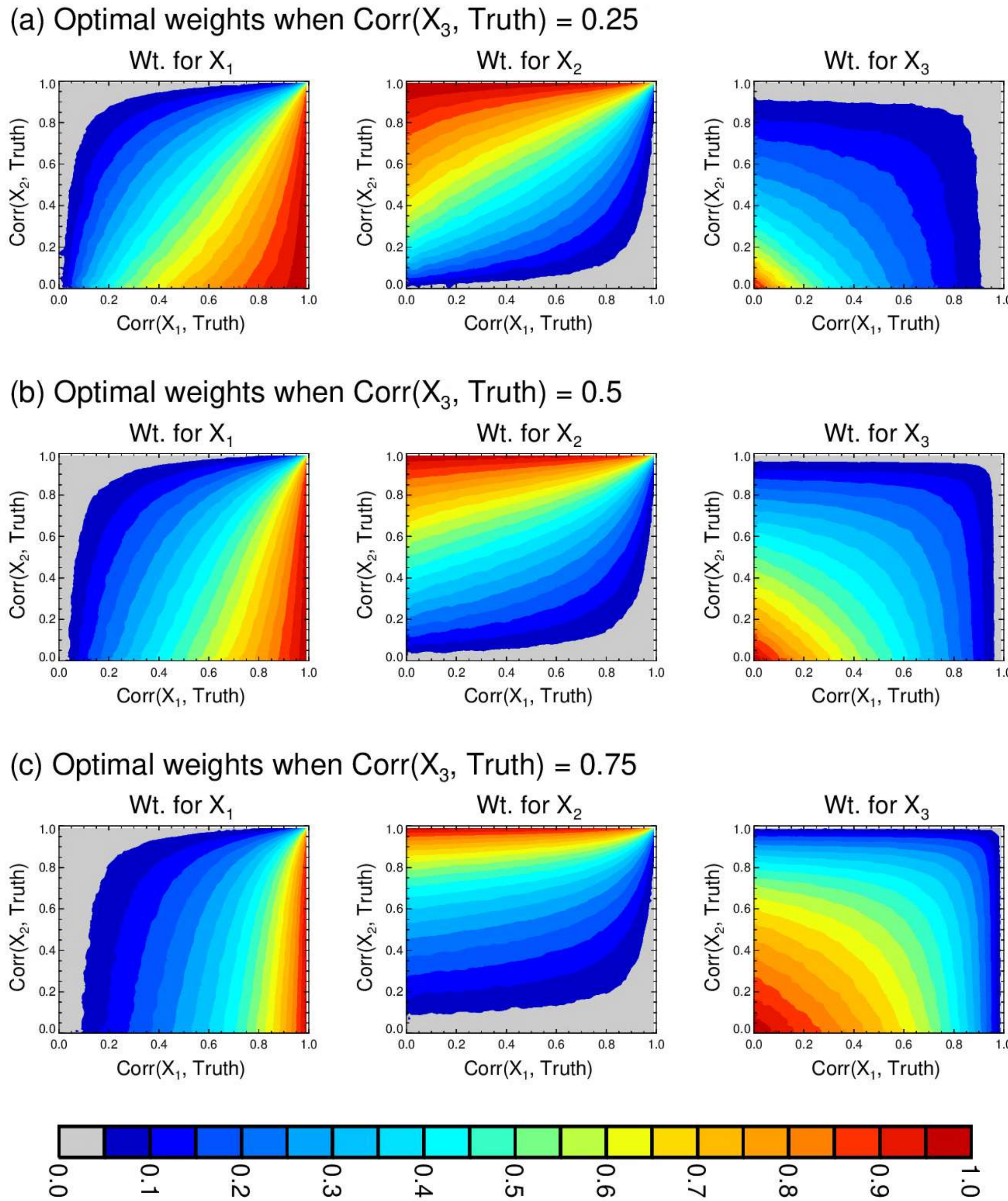

Figure 1. Optimal weights to apply to three time series (X_1, X_2, X_3) in producing a merged dataset, as a function of the correlation between each time series and the unknown truth. A full set of contours is shown for three selected values of $\rho_{3,\mathrm{Truth}}$*: (a) 0.25, (b) 0.5, and (c) 0.75.*

We apply these functions directly to the calculation of our merged precipitation data. At a given location, we first convert our logarithm time series L_G, L_I, and L_S to corresponding standard normal deviate time series Z_G, Z_I, and Z_S (for consistency with the analysis underlying Figure 1) and compute a merged time series, $Z_M(t)$:

$$Z_M(t) = W_G\, Z_G(t) + W_I\, Z_I(t) + W_S\, Z_S(t)\ . \tag{11}$$

We use the functions captured (in part) by Figure 1 to extract the weights W_G, W_I, and W_S from the values of $\rho_{G,Truth}$, $\rho_{I,Truth}$, and $\rho_{S,Truth}$ determined with (4)-(6). (Note that $\rho_{X,Truth}$ is the same for L_X and Z_X.) Using weighted mean values (μ_{ave} and σ_{ave}^2) of the means and variances of the L_G, L_I, and L_S time series, computed simply (and non-rigorously) here with

$$\mu_{ave} = W_G\, \mu_{LG} + W_I\, \mu_{LI} + W_S\, \mu_{LS} \tag{12}$$

$$\sigma_{ave}^2 = W_G\, \sigma^2_{LG} + W_I\, \sigma^2_{LI} + W_S\, \sigma^2_{LS}\ , \tag{13}$$

where μ_{LX} and σ_{LX}^2 are the mean and variance, respectively, of the logarithmic time series L_X, we expand $Z_M(t)$ into the merged precipitation estimate, $P_{merged}(t)$:

$$P_{merged}(t) = \exp(\, Z_M(t)\, \sigma_{ave} + \mu_{ave}\,)\ . \tag{14}$$

We will hereafter refer to the time series $P_{merged}(t)$ as the Merged precipitation data.

2.3 Validation Approach

2.3.1 *Global Validation Data*

We use two independent global datasets to evaluate the improvements of the Merged data over each of the contributors. The first is the time series of near-surface soil moisture retrievals provided by the Advanced Scatterometer (ASCAT) mission. ASCAT is a real aperture radar that operates at C-band; ASCAT soil moisture retrievals (reflecting moisture conditions in the top centimeter of soil) are derived from measurements of the backscatter coefficient using a semi-empirical change detection approach (Wagner et al., 2013). The processing of the ASCAT data for the present paper (version H115, from the MetOp-A and MetOp-B European Meteorological Operational spacecraft; see EUMETSAT [2019]), including the application of quality controls, is the same as that described in detail by Reichle et al. (2021); here, however, we regrid the data to the 36-km EASE grid used for the precipitation merging. An offline hydrological system (described in the next section) is used to transform the precipitation datasets into soil moisture datasets for direct evaluation against the ASCAT data.

Note that we use ASCAT data here rather than the potentially more reliable SMOS data (Kerr et al., 2010) because SMOS and SMAP data (and thus SMOS and our SM2RAIN-based precipitation data) are not adequately independent. Although the SMOS and SMAP measurements are collected from different space-borne platforms, they both use similar algorithms to convert brightness temperatures to soil moisture retrievals.

The second global dataset is the CPC near-surface air temperature (T2M) dataset, which comprises station-based T2M measurements at 0.5°×0.5° resolution (https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globaltemp.html). The data come in the form of daily minimum and maximum temperatures (T_{min} and T_{max}, respectively), which suits our purpose well, as we will be examining the correlation between precipitation and day-night temperature difference, estimated here as T_{max} - T_{min}. Prior to use, the CPC T2M data are regridded conservatively to the 36-km EASE grid.

2.3.2 *SMAP Level 4 Hydrological Modeling System.*

The four precipitation datasets (Gauge/Analysis, IMERG, SM2RAIN-based, and Merged) are each used in turn to drive the Catchment land surface model (Koster et al., 2000; Ducharne et al., 2000) globally offline on the 36-km EASE grid over the warm seasons (May through September) of 2015-2018. We use the modeling framework underlying the production of the SMAP Level 4 Version 5 product (an update of the framework underlying the Version 4 product [Reichle et al., 2019] that includes, for example, an improved aerodynamic roughness length formulation), though here we run the system without the data assimilation component. As mentioned earlier, the climatology of the Gauge/Analysis precipitation dataset is consistent with that of GPCP version 2.2; prior to running the other three precipitation datasets through the hydrological model, we scale them so that their (4-year) climatological monthly means agree with those of the Gauge/Analysis data at each grid cell. This scaling does not affect our comparisons, as we are interested here in the impacts of the short-term time variability of the precipitation fluxes rather than on the impacts of their respective climatologies. Importantly, this additional scaling makes our initialization approach more consistent with the subsequent simulation: we initialize the land model each May 1 with data from a SMAP Level 4 model-only long-term simulation on the 36-km EASE grid (i.e., a long-term simulation that uses the Gauge/Analysis data).

Note that the modeling system requires hourly precipitation data. We disaggregate the pentad values for each precipitation dataset using high temporal resolution precipitation data from the GEOS forward processing (FP) analysis system (Lucchesi, 2018) in such a way as to conserve precipitation mass. That is, for a given 5-day period and grid cell of a given precipitation dataset, the hourly precipitation values follow the sub-pentad time variability of the FP analysis, but their 5-day sum is forced to match the dataset's original pentad total. [Note that the scaling factor used is limited to a maximum of 10 but that any precipitation otherwise not included due to this cutoff is distributed within early overnight hours (midnight – 3AM local time) of the 5-day period to maintain mass conservation.] See Reichle et al. (2017a) for a description of a similar strategy applied to the use of a daily rain gauge product in a full atmospheric reanalysis.

Through these simulations, the Catchment LSM produces global fields of near-surface (0-5 cm) soil moisture across the 2015-2018 warm seasons for each precipitation dataset. We aggregate the instantaneous soil moistures to daily averages for direct comparison to the daily ASCAT data discussed above.

2.3.3 *Validation Metric.*

As noted in the introduction, we focus in this paper on evaluating the timing and relative magnitudes of the precipitation rates in the merged and contributor datasets rather than on evaluating their absolute magnitudes. For this type of evaluation, the temporal correlation (as quantified by the Pearson's correlation coefficient) against an independent dataset is the most appropriate metric, and accordingly, our validation efforts in Section 3.2 will focus on temporal correlation. This is indeed consistent with our use, in the Monte Carlo simulations underlying the construction of Figure 1, of time series correlation against an artificial truth as the target for determining optimal weights.

For our ASCAT comparisons (Section 3.2.1), we will use anomaly correlations, that is, correlations computed after the mean seasonal cycles of the time series are removed. Specifically, we will compute the square of the anomaly temporal correlation between daily ASCAT soil moisture retrievals over the warm seasons (May-September) of 2015-2018 and the corresponding soil moistures produced under each precipitation forcing. The idea is simple – because errors in the ASCAT data are completely independent of the errors in each of the contributor precipitation datasets (and completely independent of errors in the hydrological modeling system itself), higher agreement with the ASCAT data is an indication of higher precipitation accuracy. The calculation of anomaly correlations (rather than raw correlations) makes sense in the context of the ASCAT data given that the precipitation inputs to the hydrological model are already scaled to monthly climatologies (section 2.3.2), as necessitated by a need for consistency with model initial conditions.

We expect precipitation and temperature to be related for two distinct reasons: (i) the wetter soil induced by precipitation will lead (in soil moisture-limited evapotranspiration regimes) to increased evapotranspiration and thus to increased evaporative cooling, which lowers the temperature, and (ii) precipitation is associated with cloud cover, which reduces incoming solar radiation. We thus examine (Section 3.2.2) temporal correlations between the different precipitation estimates and independent T2M measurements (Section 2.3.1). For these comparisons, we have no specific reason to focus on the anomaly time series and will thus examine correlations computed on the raw time series; specifically, we will compute the square of the temporal correlation coefficient between the pentad precipitation values and the corresponding 5-day average day-night surface air temperature differences.

Note that in the figures, we will refer to the anomaly correlation between time series X and Y as anomCorr(X,Y) and to the raw correlation between them as Corr(X,Y).

3. Results

3.1 Relative Accuracy of Contributing Precipitation Datasets

We apply the extended triple collocation approach (Section 2.2) to the Gauge/Analysis, IMERG, and SM2RAIN-based precipitation data (specifically, to the logarithms of the pentad-averaged data) covering days 120-270 (roughly, May through September) of each year during 2015–2018. Equations (4)-(6) accordingly provide an estimate, at each grid cell, of the correlation between each logarithm time series and the unknown truth – that is, they provide an estimate of each dataset's inherent accuracy. Maps of the squares of these correlations (a measure of explained variance) are provided in Figures 2 and 3. Naturally, when considering these and other maps in the study, we must remember that the May through September period considered here constitutes the "warm season" for the Northern Hemisphere but the cold season for the Southern Hemisphere, with potential implications for validation.

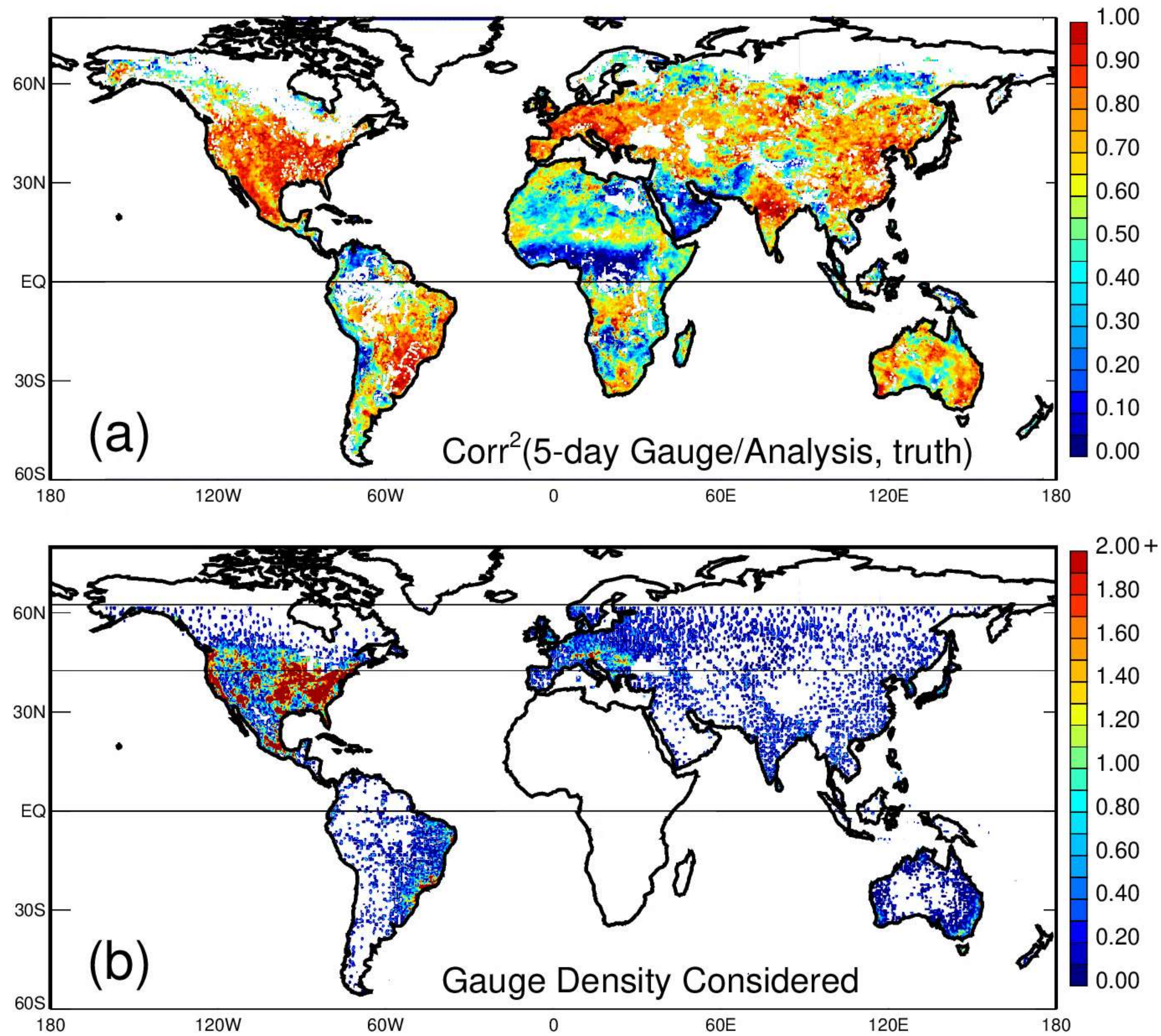

Figure 2. a. Triple collocation-based estimates of the square of the temporal correlation between the Gauge/Analysis pentad precipitation data and the unknown truth. White areas indicate where triple collocation-based estimates of accuracy were not possible given data availability (at least 100 samples from all contributors from which to compute correlations). b. Number of gauges per 0.5°x0.5° grid cell in the raw CPCU gauge-based precipitation dataset during the studied period. Data are plotted here on the 36-km EASE grid; values can be non-integers due to both the combining, through conservative regridding, of different grid cell density numbers into a single grid cell value and to the fact that the values shown represent time averages. Gauge density in Africa and north of 62.5N is not shown, as the Gauge/Analysis dataset does not utilize rain gauges in these areas (see text). The horizontal lines at 42.5N and 62.5N delimit the area over which the tapered merging of gauge data and analysis data is performed (see Reichle et al. 2017a).

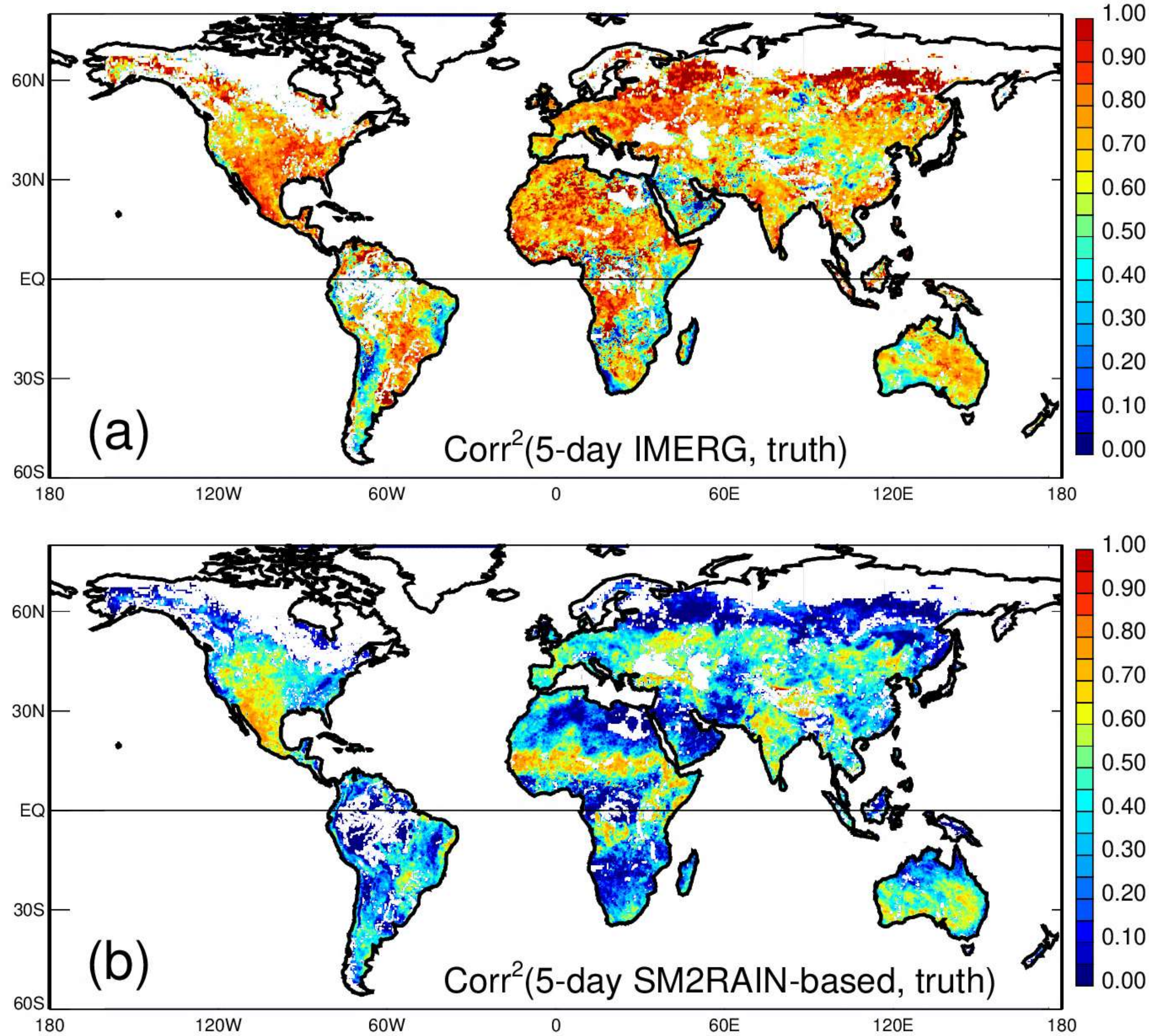

Figure 3. a. Triple collocation-based estimates of the square of the temporal correlation between the IMERG pentad precipitation data and the unknown truth. White areas indicate where triple collocation-based estimates of accuracy were not possible. b. As in (a), but for the SM2RAIN-based pentad precipitation data.

Figure 2 focuses on the Gauge/Analysis data, with Figure 2a showing the values of $\rho_{G,Truth}^2$ and Figure 2b showing the distribution of rain gauge density underlying the Gauge/Analysis dataset during the studied period. Not considering Africa and regions poleward of 62.5°N (since, as discussed in section 2.1.1, the gauge data are not used in these regions), we see that the $\rho_{G,Truth}^2$ values are clearly high only where rain gauges are present or are nearby. This, of course, is to be expected – precipitation data based on rain gauges cannot be accurate where rain gauges are not present. The comparison in Figure 2, however, is nevertheless satisfying because no explicit information regarding gauge location was used in the triple collocation analysis. The joint analysis of the three independent precipitation datasets thus effectively provides information on rain gauge density; stated another way, the consistency between Figures 2a and 2b serves as independent evidence that the triple collocation approach does provide information on the accuracy of the Gauge/Analysis dataset. Note that in regions of high gauge density, $\rho_{G,Truth}^2$ can exceed 0.9, suggesting that the gauge data in these regions do capture well the large-scale areal averages. This is consistent with the findings of Koster et al. (2019), who inferred typical length scales of precipitation correlation of hundreds of kilometers.

Figure 3 in turn focuses on the satellite-based products, with Figure 3a showing the estimated accuracy levels for the IMERG dataset and Figure 3b showing them for the SM2RAIN-based dataset. Compared to Figure 2a, the field in Figure 3a is more spatially uniform – the accuracy of the IMERG data is less variable than that of the Gauge/Analysis data across the globe. The $\rho_{S,Truth}^2$ field in Figure 3b shows that the SM2RAIN-based data are highest in semi-arid regions and are particularly low in areas with dense vegetation. This serves as further evidence that the triple collocation procedure is extracting sensible estimates of precipitation accuracy from the three precipitation datasets – SMAP retrievals are known to be inaccurate in areas of dense vegetation (Entekhabi et al., 2014), but this fact was not incorporated into the procedure. Notice that the SM2RAIN-based data show particularly low values over the very high latitudes, and they show generally lower accuracy than the IMERG data across the globe, with a few exceptions (e.g., the Sahel).

The triple collocation framework used here has an added side benefit. The weights shown in Figure 1 represent those that produce the highest degree of correlation between the merged product (in terms of logarithms) and the unknown truth for a given set of $\rho_{X,Truth}$ estimates. In establishing these weights, we identify by default this highest possible correlation. In other words, given estimates of $\rho_{G,Truth}$, $\rho_{I,Truth}$, and $\rho_{S,Truth}$ at a grid cell from (4)-(6), the analysis framework provides an estimate of the maximum level of accuracy attainable through the merging of the datasets. Figure 4 shows these maximum attainable accuracies. We see that across the globe – except in areas such as tropical forecasts and high latitudes, where at least one contributing dataset does not provide data, and in much of Africa and the Middle East, where all datasets are deficient, the potential for accuracy in the merged product is high – the square of the correlation between logarithms of the merged precipitation time series and the unknown truth generally exceeds 0.9.

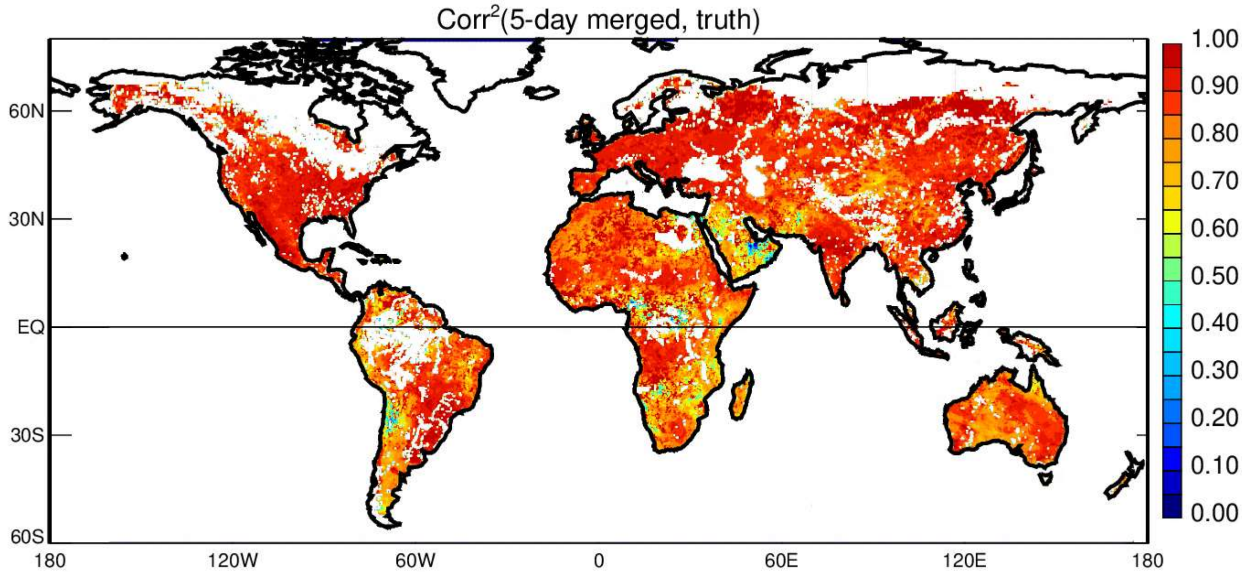

Figure 4. Triple collocation-based estimates of the maximum skill attainable from the merged precipitation dataset, expressed as the square of the temporal correlation between the merged time series and the unknown truth. White areas indicate where triple collocation-based estimates of accuracy were not possible.

Figure 5 shows the relevant differences: the idealized skill levels of the Merged dataset (Corr^2 versus unknown truth, from Figure 4) minus those for the Gauge/Analysis, IMERG, and SM2RAIN-based datasets (from Figures 2a and 3). In other words, Figure 5 shows the idealized increase in skill over each contributing dataset attainable through the merging. Figure 5 will prove especially useful for interpreting the results of our validation exercises. If the triple collocation framework is working properly, the patterns in Figure 5 should be consistent with the patterns of increased validation skill obtained with the Merged data over each contributor.

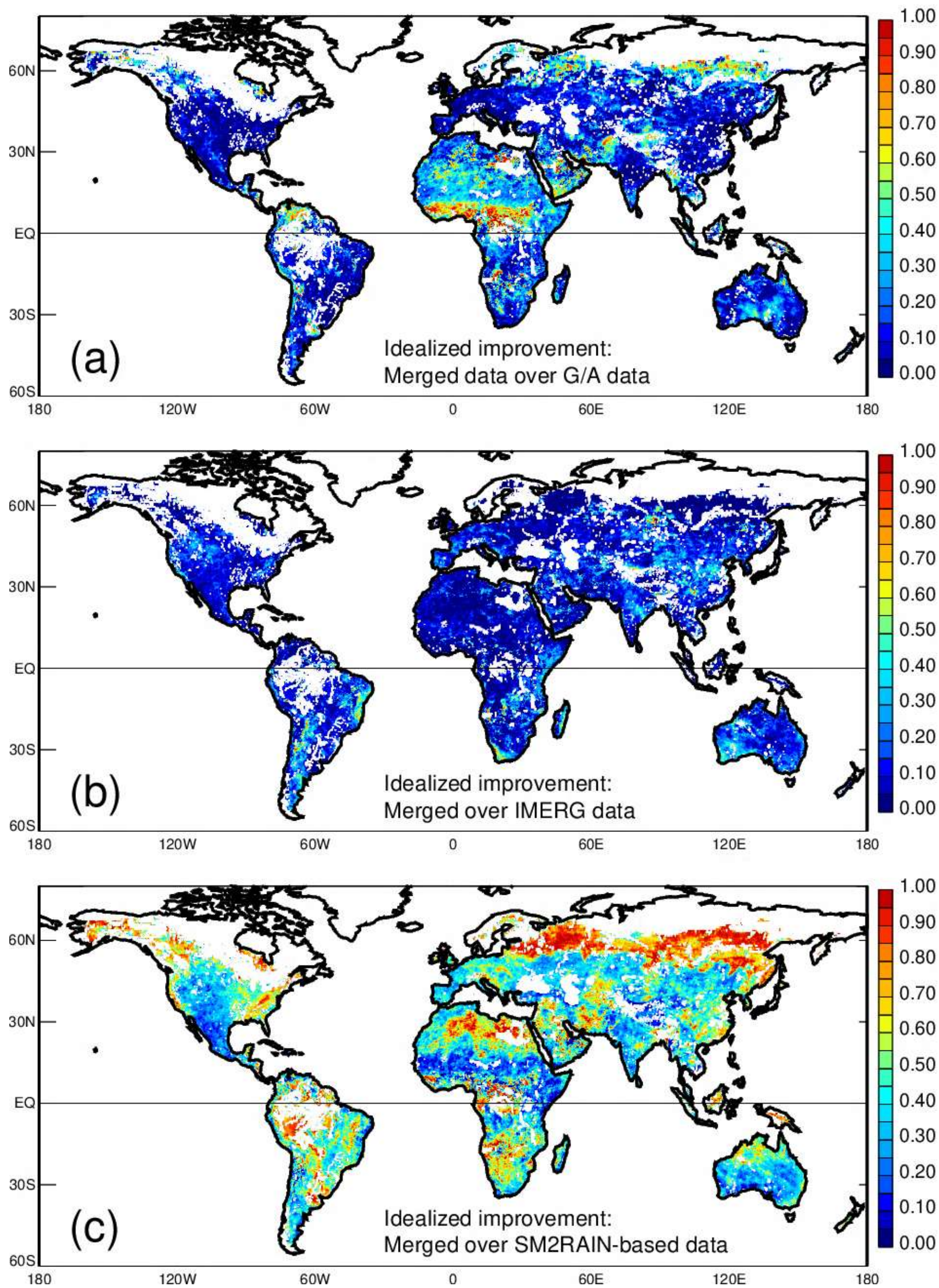

Figure 5. Degree to which the merged precipitation dataset can improve over each of the individual contributors, expressed as the difference between the maximum accuracy (square of the temporal correlation coefficient) for the Merged data shown in Figure 4 minus the accuracy estimates provided for each contributor in Figures 2 and 3. White areas indicate where triple collocation-based estimates of this improvement were not possible. a. Potential improvement of the Merged dataset over the Gauge/Analysis dataset. b. Potential improvement of the Merged dataset over the IMERG dataset. c. Potential improvement of the Merged dataset over the SM2RAIN-based dataset.

We must emphasize, however, some caveats regarding the maps in Figures 2 through 5. Even with the use of logarithms, the Gaussian assumption underlying triple collocation is violated to some degree by the presence of zero precipitation values, by the potential non-independence of Analysis and IMERG data in Africa and the high latitudes, and by the potential presence of significant seasonal cycles in the raw precipitation time series. Also, the four warm seasons of evaluation (2015-2018) provide only 120 sample pentad pairs to generate each correlation in (1)-(3), so that sampling error will affect the accuracy metrics and the subsequent quantification of the weights. (This will be discussed in more detail in Section 4.) While the consistency, for example, between the Gauge/Analysis accuracy and rain-gauge density maps in Figure 2 is encouraging, we only claim here to provide a first-order indication of the relative accuracy levels of the different precipitation products.

Approximate as they are, we use the $\rho_{G,Truth}^2$, $\rho_{I,Truth}^2$ and $\rho_{S,Truth}^2$ values shown in Figures 2 and 3 in conjunction with the functional relationships underlying Figure 1 to compute the weights used for the Merged product (Figure 6). The Gauge/Analysis data contribute the most to the Merged dataset in North America, Europe, and China, while the IMERG data contribute the most in northern Asia and Africa. The SM2RAIN-based data contribute quite a bit less than either except in a few locations (e.g., the Sahel and southern Australia). Of the three datasets, the IMERG dataset provides the most information in the global average (0.486, or 49%, as opposed to 38% for the Gauge/Analysis dataset and 13% for the SM2RAIN-based dataset). As should be expected, the Gauge/Analysis dataset provides a particularly large fraction of the information in well-gauged areas and relatively little in areas of low gauge density (Figure 2b).

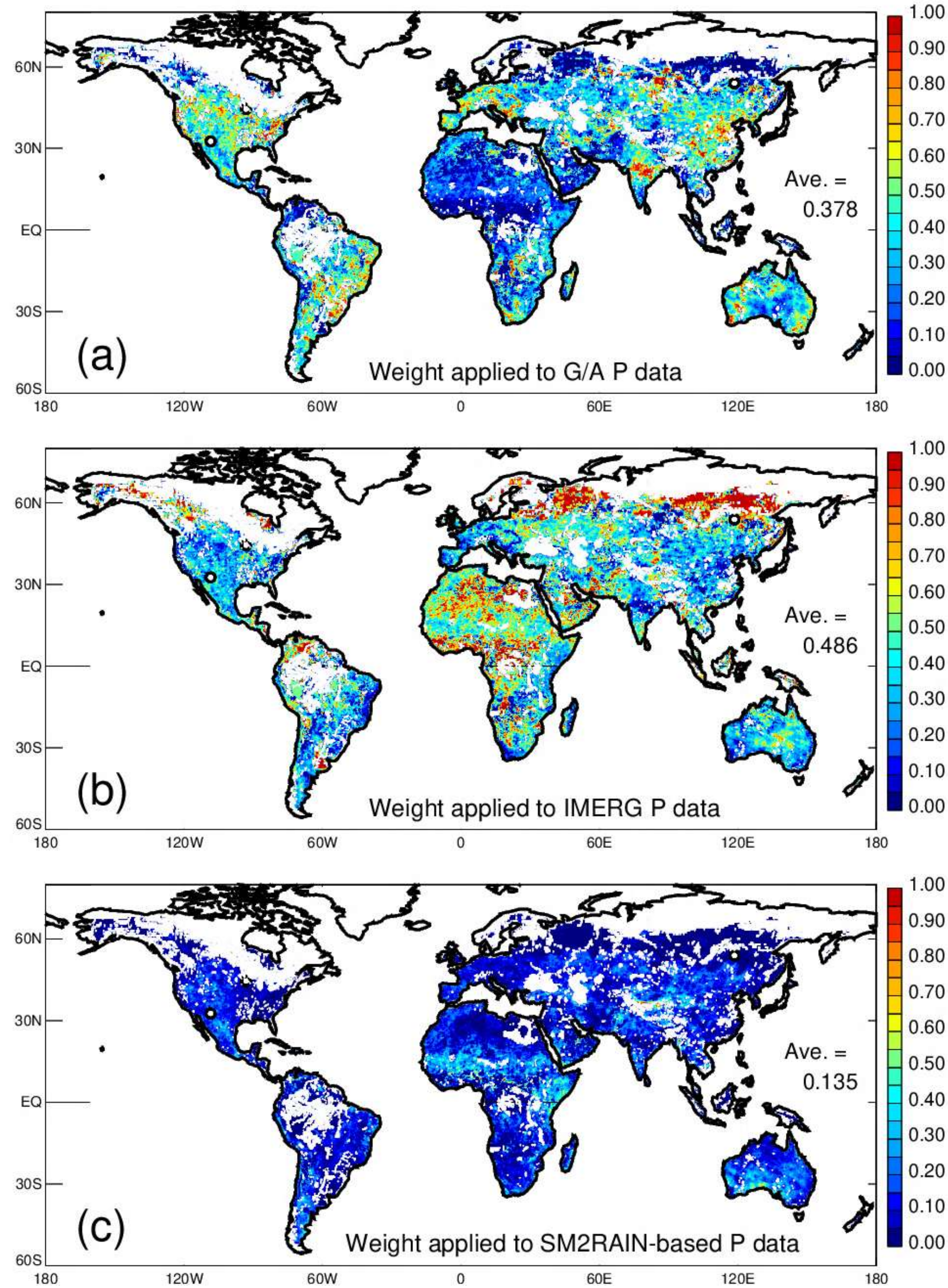

Figure 6. Weights applied in the merging process to the (a) Gauge/Analysis dataset, (b) the IMERG dataset, and (c) the SM2RAIN-based dataset. The white dots (two in North America and one in east Asia) indicate locations where sample precipitation time series will be displayed in Figure 7.

Figure 7 shows some sample time series (over 60 days during July and August 2017) to illustrate how the three datasets contribute to the Merged product. Figure 7a shows results for a grid cell in the western US; here, the weights of the three contributors are roughly the same, and indeed, the three datasets tend to agree on the timing, if not the relative magnitudes, of the precipitation events. Figure 7b shows results for a grid cell in the upper Midwest US. Here, the weight used for the Gauge/Analysis data is close to one, and accordingly, the Merged time series follows the Gauge/Analysis time series closely. Finally, Figure 7c shows results for a location in eastern Russia for which the IMERG data contributes the most to the Merged product and the Gauge/Analysis data provide a secondary correction. These two contributing time series are, in any case, quite similar during the second half of the 2-month period.

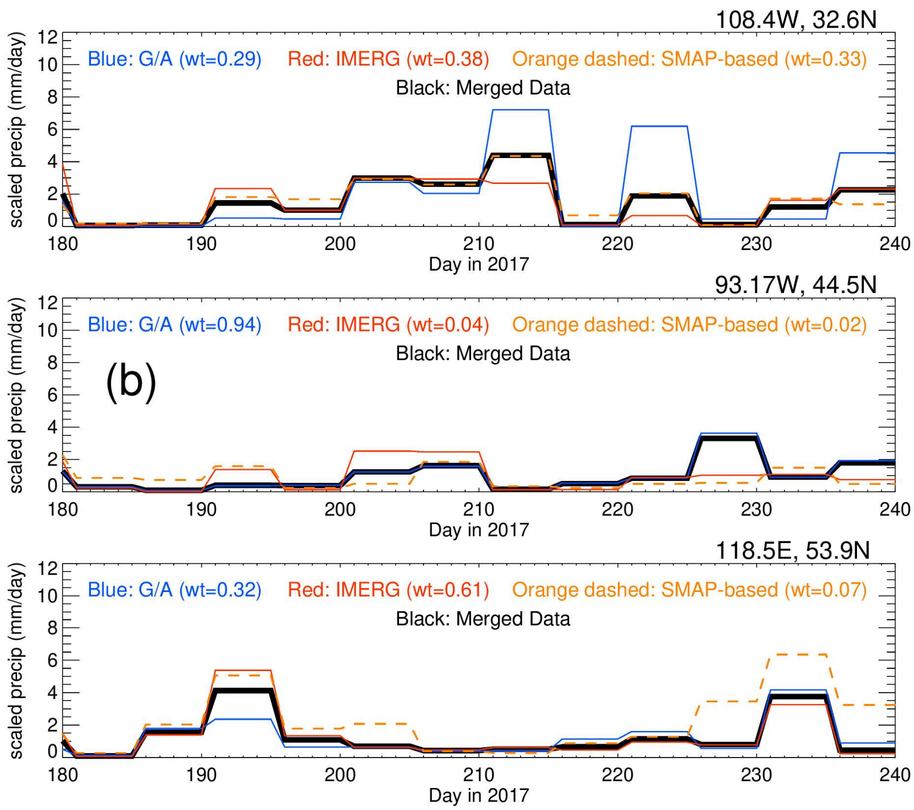

Figure 7. Sample time series of pentad precipitation rates for grid cells in: (a) the western US, (b) the upper Midwest US, and (c) eastern Russia. See Figure 6 for specific locations.

3.2 Evaluation of Products against Independent Data

Figure 5 essentially says that the Merged dataset should be inherently more accurate than each of the contributing datasets. This is not a surprise; the use of the triple collocation framework to derive the figure all but guarantees this idealized result. A much more objective evaluation of the Merged dataset requires its comparison, along with that of the contributing datasets, against wholly independent data. We provide two such comparisons in this section. We focus on comparisons against ASCAT estimates of soil moisture and against station-based observations of near-surface air temperature [Section 2.3.1].

3.2.1 ASCAT Soil Moisture Evaluations

In four separate offline simulations, we force the global hydrological model with precipitation derived in turn from the Gauge/Analysis, IMERG, SM2RAIN-based, and Merged pentad datasets, using the same sub-pentad disaggregation for each [Section 2.3.2]. The near-surface (0-5 cm) soil moisture contents generated in these simulations are now compared to ASCAT soil moisture estimates [Section 2.3.1]. While the ASCAT data are hardly free of error, these global data have the advantage of being suitably independent of the precipitation datasets being examined. If the Merged data are found to agree best with the independent ASCAT data, we take that as evidence that the Merged data are indeed more accurate than each of the three contributors.

Figure 8 shows the results in the form of difference maps: the ASCAT-based skill metric for the Merged data (the square of the anomaly temporal correlation between the ASCAT data and the soil moistures produced under the Merged precipitation forcing; see section 2.3.3) minus that for each of the three contributing precipitation datasets. Negative correlations, if they occur, are assumed to indicate a lack of skill and are set to zero before being squared. Whited-out regions in the maps have inadequate precipitation data for the triple collocation analysis or have inadequate ASCAT data for the validation exercise (Reichle et al., 2021). Positive differences in a map for a given contributor, of course, indicate that the Merged data validates better against the ASCAT data; negative differences indicate a degradation of skill.

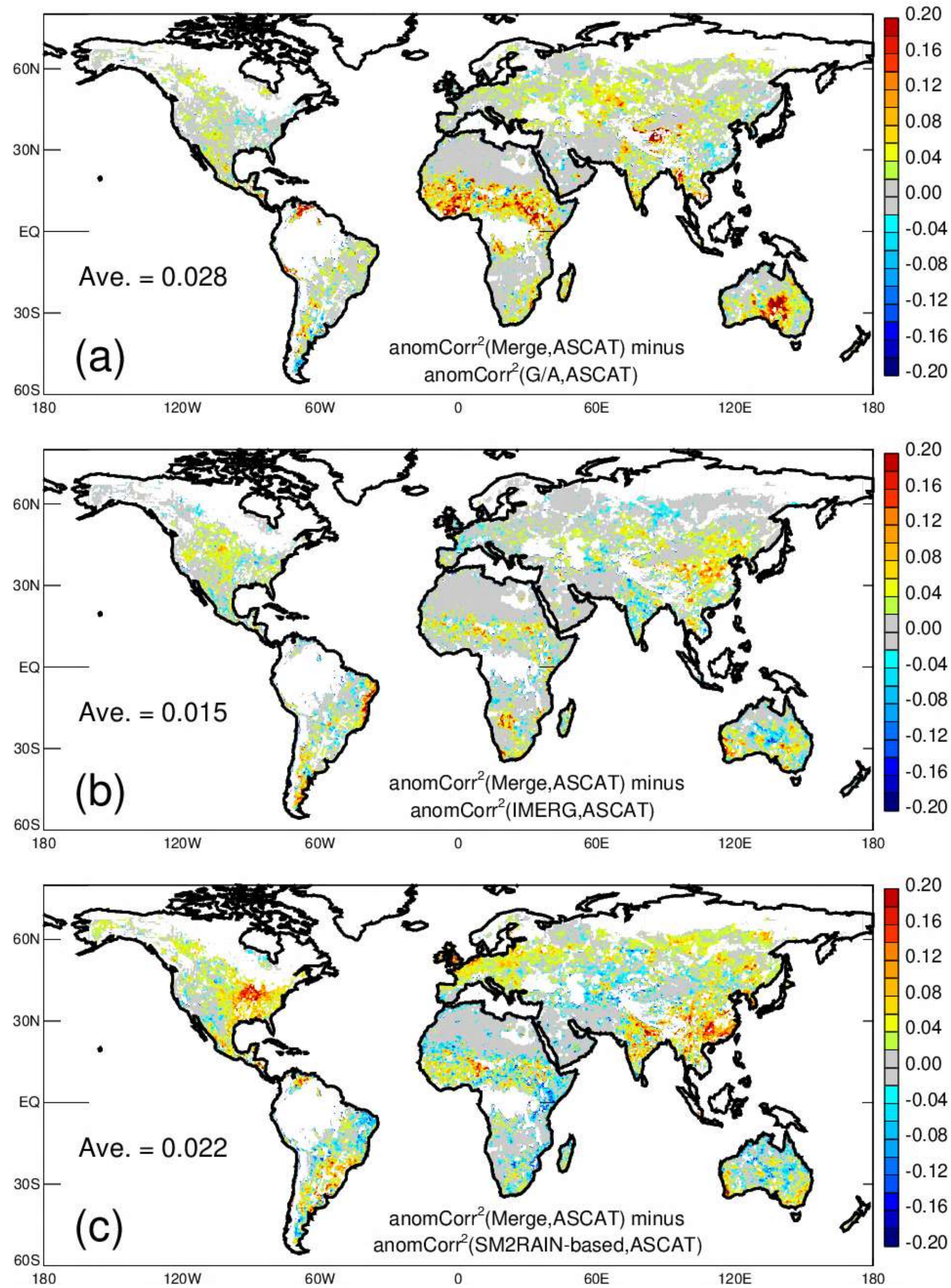

Figure 8. Degree to which the Merged dataset improves over each of the contributors when soil moistures generated with each dataset are compared to independent ASCAT measurements. (Skill is measured in terms of anomaly correlations; see Section 2.3.3.) Negative correlations are zeroed prior to squaring. White areas indicate areas for which comparisons were not possible due to limitations in the triple collocation analysis or to ASCAT data deficiencies. (a) Improvement over the Gauge/Analysis data. (b) Improvement over the IMERG data. (c) Improvement over the SM2RAIN-based data.

Notice that in all three maps, the positive values of the difference strongly outweigh the negative values – the Merged data clearly appear more accurate than any of the three contributing datasets. Furthermore, we find consistency between the difference maps in Figure 8 and the idealized difference maps in Figure 5, at least in terms of spatial patterns. Figure 5a suggests that the Merged data should perform better than the Gauge/Analysis data particularly in the Sahel, in the high latitudes of Asia, in south-central Asia, and in south-central Australia. Except for the high latitudes of Asia, this is confirmed by Figure 8a. Figure 5b indicates that the Merged data should improve over the IMERG data in, for example, the easternmost edge of South America, eastern Asia, and the southwestern corner of Australia – expectations generally supported by the validation results in Figures 8b. Finally, the expected improvements over the SM2RAIN-based data in the eastern US, South America, Europe, northern Asia, and southeast Asia (Figure 5c) also appear to a large degree in the validation results (Figures 8c). To some extent, caution is needed in comparing Figures 5 and 8, since the former is focused on total correlation and the latter on correlations computed with mean seasonal cycles removed; still, the differences in these metrics appear to be largely consistent.

While none of the contributors validate as well against the ASCAT data as do the Merged data, we note that the IMERG precipitation data appears to perform the best of the three contributors (with a global average of 0.015 for the difference metric), followed by the SM2RAIN-based data (average difference of 0.022) and the Gauge/Analysis data (average difference of 0.028). The fact that the SM2RAIN-based data appears to perform better than the Gauge/Analysis data here is curious, given the opposite expectation indicated in Figure 5. We have no clear explanation for this inconsistency, other than to say that the SM2RAIN-based precipitation dataset is derived from soil moisture retrievals, which may potentially give it some advantage in a comparison focused on soil moisture. Also, ASCAT data limitations eliminated from consideration in Figure 8 some regions (e.g., parts of the Amazon and the Congo) for which improvements of the Merged data over the SM2RAIN-based data were expected from Figure 5c to be particularly large. Presumably, if we had been able to consider those regions in the ASCAT analysis, the global average of the difference metric for the SM2RAIN-based data would have been somewhat larger.

With the Instrumental Variable approach (Su et al., 2014), the accuracy of soil moisture data can be quantified by analyzing it in conjunction with ASCAT data at different time lags (Reichle et al., 2021). When we apply this approach (which itself is based on triple collocation) to our data here, the resulting relative skill levels associated with the different precipitation datasets are highly consistent with those shown in Figure 8 – the Merged data improves over the individual datasets to the same relative degree, following basically the same spatial patterns. The Instrumental Variable results are provided as Figures S1-S3 in the Supporting Information.

3.2.2 Station-Based Air Temperature Evaluations

As noted in Section 2.3.3, a negative temporal correlation between precipitation and air temperature (T2M) can be expected given that wetter soils induce greater evaporative cooling and because precipitation periods are associated with increased cloudiness, which reduces the incoming solar radiation. Both mechanisms particularly promote a negative correlation between precipitation and the day-night temperature difference. We examine here the degree to which the precipitation datasets on their own (not run through the hydrological model) capture such a negative correlation.

Of course, a comparison like this is fraught with caveats, since temperature variations depend on other factors as well – advection of warm or cool air masses, impacts of aerosols and non-precipitating clouds on the radiation balance, and so on. Perhaps more importantly, the coverage of station temperature measurements is far from uniform across the globe. As will be seen, this latter issue may, in some regions, limit the usefulness of the comparisons of the Merged data results against those of the three contributors.

With these caveats in mind, we present in Figure 9 the relevant differences: the accuracy metric for the Merged dataset (the square of the temporal correlation between the pentad precipitation values and the corresponding 5-day average day-night surface air temperature differences; see section 2.3.3) minus that for each of the contributing datasets. Positive correlations imply that the aforementioned mechanisms do not manifest themselves in the data and are thus zeroed before they are squared. Overall, the results look promising. In the temperature-based evaluation, the Merged data improve over the Gauge/Analysis data (Figure 9a) in many of the same regions indicated in the ASCAT-based evaluation: the Sahel, parts of central Australia, and south-central Asia. The global mean difference is positive (0.010). Setting aside momentarily the skill degradation in northern Asia (discussed further below), the general consistency between the ASCAT-based results (Figure 8a) and the temperature-based results (Figure 9a) supports the idea that the improvements shown are real, as does the comparison of the temperature-based results against the idealized improvements in Figure 5a.

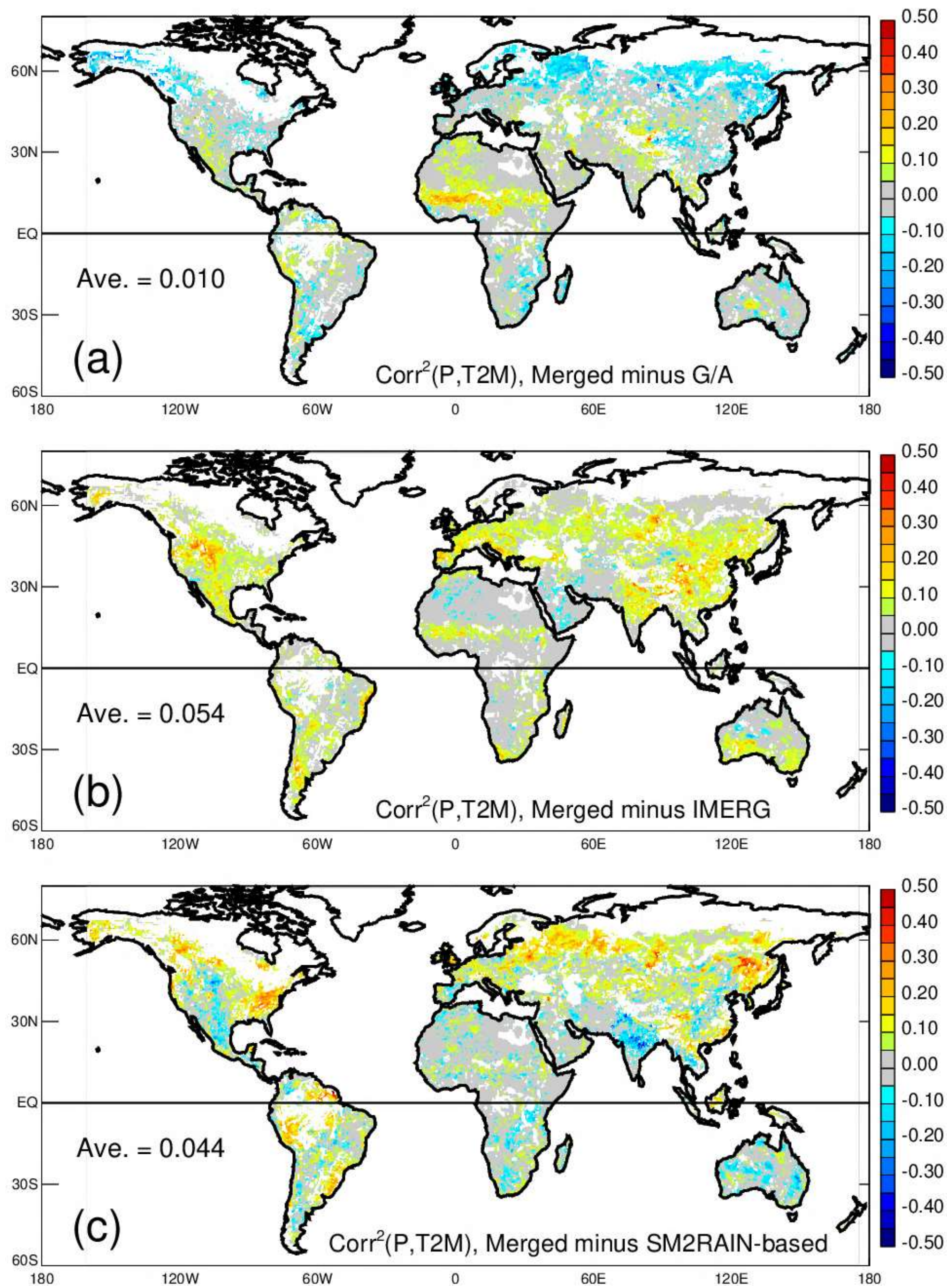

Figure 9. Degree to which the Merged dataset improves over each of the contributors when each dataset is correlated against gridded near-surface air temperature (T2M) differences (daily maximum T2M minus daily minimum T2M). Negative correlations are expected, so positive correlations are zeroed prior to squaring. White areas indicate areas for which comparisons were not possible due to limitations in the triple collocation analysis. (a) Improvement over the Gauge/Analysis data. (b) Improvement over the IMERG data. (c) Improvement over the SM2RAIN-based data.

Figures 9b and 9c show an even stronger improvement of the Merged product over the IMERG and SM2RAIN-based products, respectively, using this metric. Here, for reasons discussed further below, we emphasize not the magnitudes of these improvements but the general agreement between their spatial patterns and those seen for both the corresponding ASCAT-based results and the idealized analysis in Figures 5b and 5c. For example, the T2M-based results and idealized differences both show (as did the ASCAT-based results) a general improvement of the Merged data over the IMERG data in the neighborhood of Montana in the US as well as in China and the southwest corner of Australia. The skill differences for the SM2RAIN-based data are large in the eastern US for both the T2M-based and idealized results (again, as they were for the ASCAT results), and similar agreement also appears over northern Asia.

Some features of the T2M-based results, however, are not easily explained. Again, the accuracy metric for the Merged dataset is consistently lower than that for the Gauge/Analysis data along the northern reaches of Asia (though keep in mind that the blue areas are artificially amplified here by the Mercator projection). We can speculate on a partial explanation. Given that very few stations in this area measure T2M (see, e.g., Figure 2 of Fan and van den Dool) [2008]), the temperature measurements underlying the calculation in many grid cells are largely based on individual point measurements, often from points lying outside the grid cell in question, and these point measurements are likely coincident with the point precipitation measurements contributing to the Gauge/Analysis data. That is, in the construction of the global gridded air temperature and precipitation products, the data provided for many grid cells in northern Asia may – in effect – consist of collocated point measurements of precipitation and air temperature from a single remote location. As a result, because the correlation between precipitation and air temperature may very well be strong at point measurement sites, the representativeness error noted above for precipitation gauge data does not manifest itself as a degradation in our computed correlations.

The Gauge/Analysis data thus have an advantage over the IMERG and SM2RAIN-based data: for the temperature-based evaluation, a chief source of error in the Gauge/Analysis data does not limit the performance of these data. We speculate that if the temperature data themselves were truly representative of local grid cell averages (rather than point values at potentially remote measurement stations) in the northern reaches of Asia, the Merged data might indeed appear more accurate there. Again, though, this would only be a partial explanation, as the rain gauges contribute less and less information to the Gauge/Analysis product as one approaches 62.5N and none at all north of that latitude (see section 2.1.1). Some other feature of high northern latitude meteorology (e.g., weather that is more strongly dominated by advection or a snow season with a later end date or an earlier onset date) may be muddying our analysis in this area; also, strong seasonal cycles in the high latitudes may have adversely affected our triple collocation analysis (see Section 2.2).

This spatial representativeness argument may also help explain why the improvement of the Merged data over either the IMERG data or the SM2RAIN-based data in Figure 9 is so much larger than the improvement over the Gauge/Analysis data. If a low density of both precipitation and T2M measurements does indeed allow the correlation calculation for the Gauge/Analysis data to bypass the ill effects of spatial representativeness error, this benefit will also be transferred preferentially to the Merged data, to the extent that the latter are derived from the former. In other words, the improvements seen in Figures 9b and 9c are probably somewhat exaggerated in regions of low measurement density. Such arguments, however, would not explain why the improvement of the Merged data over the IMERG data is somewhat higher than that over the SM2RAIN-based data in the global average, in contradiction to expectations (Figure 5). The SM2RAIN-based data's correlations with T2M are particularly better than those of the IMERG data in the central US, India, southern Africa, and Australia. Perhaps this is related to the fact that one of the two mechanisms underlying the expected correlation directly involves soil moisture, information that is directly built into the SM2RAIN-based data.

For these reasons, and because temperature measurements are in fact absent in many parts of the globe (e.g., over much of the Southern Hemisphere), the CPC T2M analysis above arguably pales to that of our earlier ASCAT analysis as a means of evaluating the four precipitation datasets. Even so, the T2M-based evaluation – particularly the consistency in the spatial patterns with the ASCAT-based evaluation and with the idealized differences in Figure 5 – generally supports the idea that on a global scale, the Merged product is more accurate than each of the three contributors.

4. Discussion

The three precipitation datasets contributing to the Merged dataset have independent errors. Thus, to the degree that these datasets satisfy the other requirements of triple collocation (when processed as described, using logarithms of pentad totals), the Merged dataset should, at least according to theory, capture the time variability of the pentad rainfall better than any of the three contributors individually. Our ability to illustrate this conclusively, however, is necessarily limited in two important ways.

First, fully independent data are required for the validations, and, at least on the global scale, such data are rare. Our global-scale evaluations in Section 3.2 are accordingly limited to the use of ASCAT soil moisture data and CPC air temperature data, with evaluations against the latter being particularly indirect. Fortunately, these evaluations prove, on balance, to be successful, even if the Merged data do not perform better in every location on the globe.

Local-scale evaluations against in-situ measurements could, of course, be used to supplement the global-scale evaluations. In the course of our work, we compared the output of our four global hydrological simulations against in-situ soil moisture measurements – the same measurements Reichle et al. (2019) used to validate SMAP Level 4 products. We relegate these results to the Supporting Information (see Figures S4 and S5) because they mainly reflect the fact that most of these in-situ measurements were taken over the continental US, a region for which the Merged dataset overwhelmingly reflects the Gauge/Analysis dataset (see weights in Figure 6). Accordingly, these comparisons fully agree with expectations: when validated against in-situ soil moisture data, the Merged dataset performs better (sometimes significantly so) than the IMERG and SM2RAIN datasets and only slightly better than the Gauge/Analysis dataset. Across the globe, in-situ hydrometeorological measurements of quality and duration suitable for validation indeed tend to be taken in areas that also feature high rain gauge coverage, i.e., locations for which it would be difficult to illustrate conclusively the advantages of the Merged dataset. This is a common limitation of such local-scale evaluations.

The second important obstacle to the evaluation of our merging approach has to do with the length of the observational record and the associated uncertainty in the weights assigned to the three contributors. Of the six years of SMAP data available to us, two were used to calibrate the SM2RAIN-based algorithm, leaving four years of data to use in the merging – a total of only

120 warm-season (May-September) pentads. As a result, the correlations in (1)-(3) will suffer from sampling error, and this error will be compounded when ratios of the correlations are taken in (4)-(6) to compute $\rho_{G,Truth}$, $\rho_{I,Truth}$, and $\rho_{S,Truth}$.

To investigate the potential impacts of sample size, we now consider three measurement time series (X_1, X_2, X_3) with independent errors and with known (prescribed) correlations against the time series, Truth(t), of actual values. (We thus follow here the idealized framework underlying Figure 1.) Using the known values of $\rho_{1,Truth}$, $\rho_{2,Truth}$, and $\rho_{3,Truth}$, we first use (7)-(9) to construct sample sets of measurement time series of a given length. These time series are then used in turn to estimate, with sampling error: (i) the correlations in (1)-(3), (ii) corresponding correlations against truth using (4)-(6), and (iii) the resulting weights to apply to each dataset using the algorithm underlying Figure 1. Finally, we apply (7)-(9) to generate lengthy versions of X_1, X_2, and X_3 (i.e., time series long enough so that sampling error is not an issue) and use the imperfect weights generated in step (iii) above to generate a lengthy, but imperfect, time series of the Merged data, which we then correlate against the known truth. The process is repeated 1000 times to obtain an average correlation against truth for the imperfect Merged time series.

Figure 10 shows the results – for four different underlying sample sizes, it provides the average correlation against truth as a function of $\rho_{1,Truth}$ and $\rho_{2,Truth}$. (For illustration purposes, $\rho_{3,Truth}$ is taken to be 0.5 throughout.) The upper left panel shows the average accuracies achieved with the Merged dataset when the sample size (time series length) underlying the estimation of the weights is 120, as it was for our calculations in Section 3. The next two panels show results for sample sizes of 500 and 1000, respectively. The lower right panel shows the accuracies obtained when the weight estimation is not limited by time series length.

Clearly, longer sample sizes bring the estimated accuracies closer to the optimal values in the lower right panel of the figure. Notice, however, that even with a sample size of 120, substantial accuracy is still achieved. Consider the example indicated by the small white dot in the panels, which represents the case for which $\rho_{1,\mathrm{Truth}}$, $\rho_{2,\mathrm{Truth}}$, and $\rho_{3,\mathrm{Truth}}$ are all equal to 0.5. The lower right plot shows that with no sampling error, the average correlation of the Merged data against truth in this case would be about 0.71. When we construct the Merged data with weights made sub-optimal by sampling error, this correlation does go down, but even with a sample size of 120 underlying the weights, the correlation of the Merged data against truth reduces to only 0.66, which is still well above 0.5, the correlation of each of contributor against truth. Examples like this give us confidence that our merging process can be effective, even with such short time series lengths.

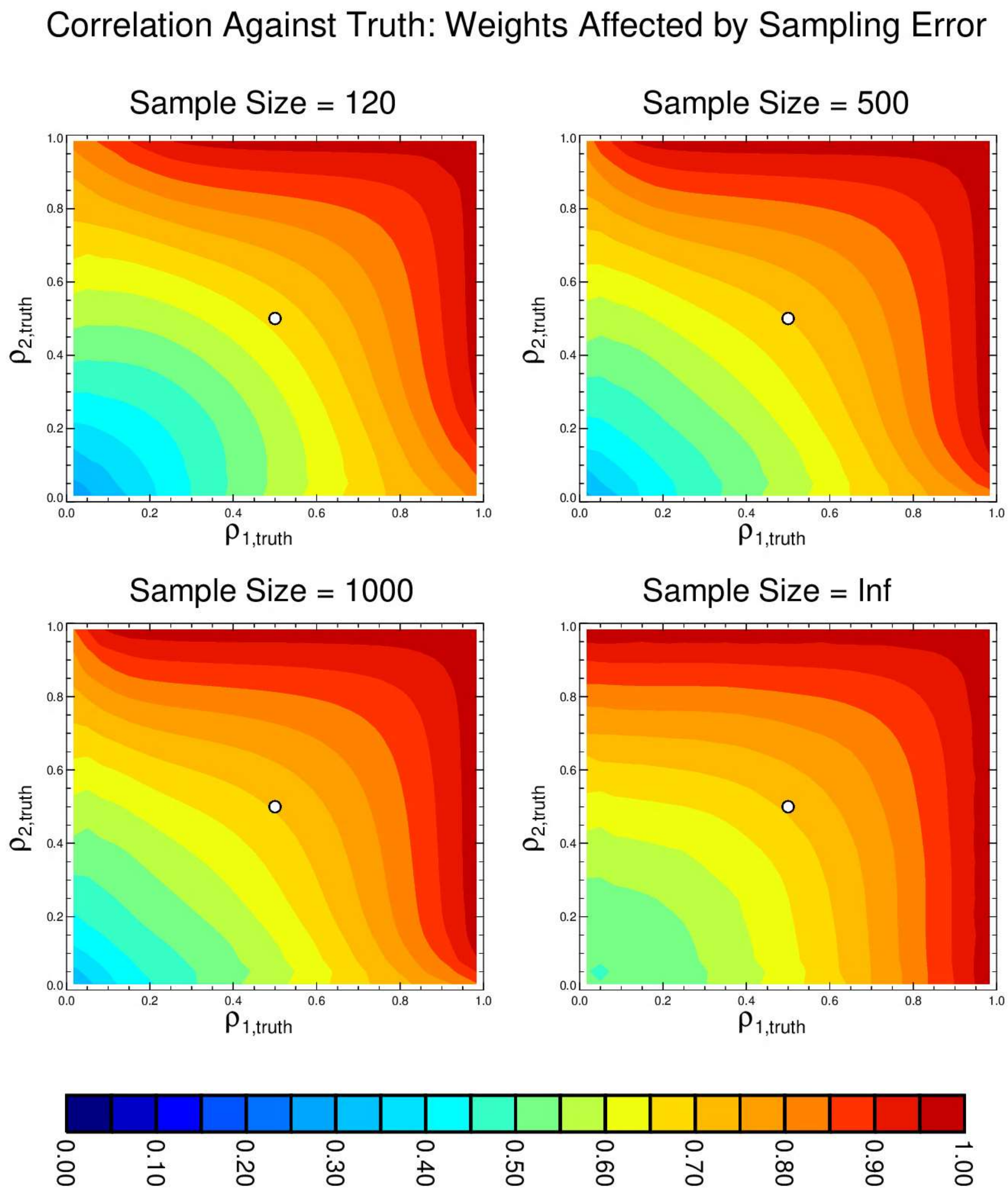

Figure 10. Impact of sample size on the effectiveness of the merging procedure, as revealed by an idealized Monte Carlo analysis (see text). Shown are the average correlations against truth of the merged data as functions of $\rho_{1,\text{Truth}}$ and $\rho_{2,\text{Truth}}$ *(i.e., the prescribed correlations between time series X_1 and X_2, respectively, and the unknown truth);* $\rho_{3,\text{Truth}}$ for the third time series, X_3, is set to 0.5 for all plots. *Top left: results for the case when the weights for the merging are determined from time series with length 120. Top right: Same, but for weights based on time series of length 500. Bottom left: Same, but for weights based on time series of length 1000. Bottom right: Same, but for weights based on (effectively) infinite time series length. The example indicated by the small white dot is discussed in the text.*

Note that up to this point in our study, we have not considered alternative data fusion methods. Again, theory suggests that if the triple collocation assumptions are satisfied (regarding the normality of the underlying distributions, the constancy of the statistics describing these distributions, and so on), the weights we derive with our approach (assuming sufficiently long time series; see above) should indeed be optimal. As already discussed, however, the assumptions underlying triple collocation are not perfectly satisfied in the real world, and thus the merged data are not guaranteed to be closer to the truth than each contributor. One might reasonably ask if an alternative approach to deriving the weights – an approach not limited by these assumptions – could perform better. Side-by-side analysis of the triple collocation approach with alternative advanced data fusion methods [e.g., that of Beck et al. (2017) or one involving, for example, Kalman filtering or machine learning] is beyond the scope of this study. However, it is straightforward to test our approach against a simple "consensus" approach, one in which all contributors are given equal weight in the merging. Such simple averaging has proven effective in past studies. Fritsch et al. (2000), for example, showed that consensus short-term weather forecasts constructed from forecasts produced by multiple systems proved superior to the individual contributing forecasts.

We thus constructed a consensus precipitation product by applying equal weights (1/3) to the Gauge/Analysis, IMERG, and SM2RAIN-based precipitation products, and we then used this consensus product to drive the hydrological model. The degrees to which the two merging approaches generate soil moistures that agree with the ASCAT observations (as measured with the squared anomaly correlation metric) are compared with the difference map in Figure 11, with positive differences indicating that the original (triple collocation-based) Merged dataset performs better than the simple consensus dataset. Some negative differences appear in the map, but overall, the map is dominated by positive differences. The fact that the global average of the differences (0.0063) is smaller than the globally-averaged differences seen in Figure 8 (note the reduced range on the color bar in Figure 11) suggests that the simple consensus averaging approach does extract complementary skill from the three contributing datasets. Importantly, though, the positive value of this difference (smaller, but of the same order as the averaged differences seen in Figure 8) supports the idea that our original merging approach produces weights that are indeed more optimal. Perhaps, if the underlying time series were longer and the triple collocation-based weights were accordingly more accurate, the improvement indicated in Figure 11 would be even more extensive.

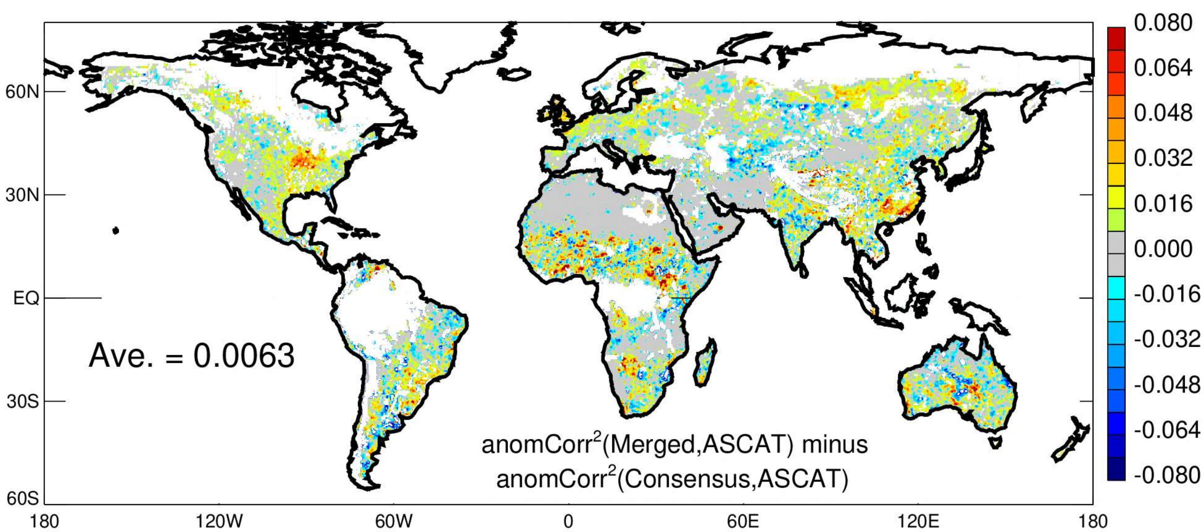

Figure 11. As in Figure 8, but showing the degree to which the Merged dataset improves over a "consensus" merging of the Gauge/Analysis, IMERG, and SM2RAIN-based data in which each contributor is assigned equal weight (0.33). White areas indicate areas for which comparisons were not possible due to limitations in the triple collocation analysis or to ASCAT data deficiencies.

5. Summary

In this study, three fully independent 36-km, pentad precipitation datasets (Gauge/Analysis, IMERG, and SM2RAIN-based) were examined together in a triple collocation framework. The analysis provides estimates of the skill (square of the temporal correlation) of each dataset against the unknown truth (Figures 2 and 3). Given the limited sample size and limitations in satisfying certain triple collocation assumptions, these estimates represent, at best, first-order estimates of what is otherwise an unmeasurable property. Even so, it is encouraging that the quantified skill distributions are broadly consistent with, for example, rain gauge density distributions and known limitations in SMAP retrievals.

Using these estimates of inherent dataset skill, we combined the three pentad datasets into a single Merged product, applying weights that optimize the expected correlation between the merged product and the unknown truth. In theory, this Merged dataset takes advantage of the particular strengths of each contributor and accordingly should be more accurate than each on its own. To test this, we evaluated the relative accuracy of the Merged product and the three contributor datasets against two separate and fully independent global datasets: ASCAT soil moisture retrievals and station-based T2M measurements. The Merged product clearly performs better than each of the contributors in the ASCAT comparisons (which involve output generated with a global offline model forced with each of the precipitation datasets). The T2M comparisons are inherently more limited; even so, the Merged product again shows improved performance relative to each of the contributors. Furthermore, the patterns in the improved performance are generally consistent with expectations from the triple collocation framework (Figure 5). We thus conclude that the Merged data are in fact generally more accurate on a global scale than any of the three contributors, having taken advantage of the relative strengths of each.

The generation of an improved pentad precipitation dataset should not be considered an end in itself. The present work demonstrates that at a spatial scale of 36-km and a temporal scale of 5 days, the merged pentad product does take advantage of the strengths of each contributor. The raw versions of the contributors, however, provide information at higher spatial and temporal resolutions. IMERG provides particularly high resolutions: half-hourly data at scales of 10-km. An obvious next step is to disaggregate the optimized Merged pentad data using sub-pentad, sub-36-km resolution precipitation information contained in, for example, the IMERG dataset to produce data that might be even more effective for hydrological simulation.

Acknowledgments. Funding for this work was provided by the NASA SMAP mission and the SMAP Science Team. Computational resources were provided by the NASA High-End Computing program through the NASA Center for Climate Simulation. We thank Sarith Mahanama, David Bolvin, and Yehui Chang for help with the datasets.

Data Availability Statement. The SMAP L2 retrievals underlying the SM2RAIN-based data are available from http://nsidc.org/data/smap. IMERG data are archived at NASA GES DISC (https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_06/summary). GEOS forcing data are available from https://fluid.nccs.nasa.gov/weather. CPCU precipitation is available from ftp://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_CONUS, and CPC Global Temperature data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, at https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globaltemp.html. ASCAT soil moisture data are available from the European Organisation for the Exploitation of Meteorological Satellites Hydrology Satellite Application Facility at http://hsaf.meteoam.it/soil-moisture.php. The GEOS source code is available under the NASA Open-Source Agreement at http://opensource.gsfc.nasa.gov/projects/GEOS-5.

References

Adler, R.F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., et al. (2003). The Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-Present). *J. Hydrometeor.*, *4*,1147-1167.

Arkin, P. A., & Meisner, B. N. (1987). The relationship between largescale convective rainfall and cold cloud over the Western Hemisphere during 1982–84. *Mon. Wea. Rev*., *115*, 51–74. doi 10.1175/1520-0493(1987)115<0051:TRBLSC>2.0.CO;2

Arnaud, P., Lavabre, J., Fouchier, C., Diss, S., & Javelle, P. (2010). Sensitivity of hydrological models to uncertainty in rainfall input. *Hydrol. Sci. J*., *56*, 397-410. doi: 10.1080/02626667.2011.563742

Beck, H.E., van Dijk., A.I.J.M., Levizzani, V., Schellekens, J., Miralles, D.G., Martens, B., & de Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.*, *21*, 589–615. doi: 10.5194/hess-21-589-2017

Bisselink, B., Zambrano-Bigiarini, M., Burek, P., & de Roo, A. (2016). Assessing the role of uncertain precipitation estimates on the robustness of hydrological model parameters under highly variable climate conditions. *J. Hydro. Reg. Studies*, *8*, 112-129. doi:10.1016/j.ejrh.2016.09.003.

Bras, R., and Rodriguez-Iturbe, I (1985). Random Functions and Hydrology. Addison Wesley, Dover Publications, New York.

Brocca, L., Moramarco, T., Melone, F., & Wagner, W. (2013). A new method for rainfall estimation through soil moisture observations. *Geophys. Res. Lett*., *40*, 853-858.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., et al. (2014). Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data, *J. Geophys. Res. Atmos.*, *119*, 5128-5141, doi:10.1002/2014JD021489.

Brodzik, M. J., Billingsley, B., Haran, T., B. Raun, & Savoie, M. H. (2012). EASE-Grid 2.0: Incremental but significant improvements for Earth-gridded data sets. *ISPRS International Journal of Geo-Information*, *1(1)*, 32-45.

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

Chiaravalloti, F., Brocca, L., Procopio, A., Massari, C., & Gabriele, S. (2018). Assessment of GPM and SM2RAIN-ASCAT rainfall products over complex terrain in southern Italy. *Atmospheric Research*, *206*, 64-74. doi:10.1016/j.atmosres.2018.02.019.

Ciabatta, L., Massari, C., Brocca, L., Gruber, A., Reimer, C., Hahn, S., et al. (2018). SM2RAIN-CCI: a new global long-term rainfall data set derived from ESA CCI soil moisture. *Earth System Science Data*, *10*, 267-280. doi:10.5194/essd-10-267-2018.

Dong, J., Lei, F., & Wei, L. (2020). Triple collocation based multi-source precipitation merging. *Frontiers in Water*, *2*, doi:10.3389/frwa.2020.00001.

Draper, C., Reichle, R., de Jeu, R., Naeimi, V., Parinussa, R., & Wagner, W. (2013): Estimating root mean square errors in remotely sensed soil moisture over continental scale domains. Remote Sens. Env., 137, 288-298. https://doi.org/10.1016/j.rse.2013.06.013

Ducharne, A, Koster, R. D., Suarez, M. J., Stieglitz, M., & P. Kumar, P. (2000). A catchment-based approach to modeling land surface processes in a general circulation model. (2) Parameter estimation and model demonstration. *J. Geophys. Res. 105*(D20), 24823–38.

Entekhabi, D., Njoku, E., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The Soil Moisture Active Passive (SMAP) mission. *Proc. IEEE*, *98*, 704-716.

Entekhabi, D., Yueh, S., O'Neill, P. E., Kellogg, K. H., Allen, A., Bindlish, R., et al. (2014). SMAP Handbook, Mapping Soil Moisture and Freeze/Thaw from Space. SMAP Project, JPL 400-1567, National Aeronautics and Space Administration, Pasadena, CA.

EUMETSAT (2019). Product User Manual (PUM) Metop ASCAT Soil Moisture Climate Data Record v5 12.5 km (H115) and extension (H116). *EUMETSAT Satellite Application Facility on Support to Operational Hydrology and Water Management*, Doc. SAF/HSAF/CDOP3/PUM/, Revision: 0.1, 32 pp., https://hsaf.meteoam.it/documents/H115_ASCAT_SSM_CDR_v5_PUM_v0.1.pdf.

Fan, Y. & van den Dool, H. (2008). A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, *113*, D01103. doi:10.1029/2007JD008470.

Fritsch, J. M., Hilliker, J. Ross, J., & Vislocky, R. L. (2000). Model consensus. Weather and Forecasting, 15, 571-582, https://doi.org/10.1175/1520-0434(2000)015<0571:MC>2.0.CO;2 .

Huffman, G.J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Kidd, C., et al. (2020). Integrated Multi-satellitE Retrievals for the Global Precipitation Measurement (GPM) mission (IMERG). Chapter 19 in *Adv. Global Change Res*., Vol. *67*, Satellite Precipitation Measurement, V. Levizzani, C. Kidd, D. Kirschbaum, C. Kummerow, K. Nakamura, F.J. Turk (Ed.), Springer Nature, Dordrecht, ISBN 978-3-030-24567-2 / 978-3-030-24568-9 (eBook), 343-353. doi:10.1007/978-3-030-24568-9_19

Kedem, B. & Chiu, L. S. (1987). On the lognormality of rain rate. Proc. National Acad. Scie., 84, 901-905, doi: 10.1707/pnas.84.4.901.

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., et al. (2010). The SMOS mission: New tool for monitoring key elements of the global water cycle. *Proc. IEEE*, *98*, doi:10.1109/JPROC.2010.2043032.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., & Kirschbaum, D. B. (2017). So, how much of the Earth's surface is covered by rain gauges? *Bull. Amer. Meteor. Soc*., *98*, 69-78. doi:10.1175/BAMS-D-14-00283.1

Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M., & Kumar, P. (2000). A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure, *J. Geophys. Res*., *105(20)*, 24,809–24,822.

Koster, R. D., Brocca, L., Crow, W. T., Burkin, M. S., & De Lannoy, G. J. M. (2016). Precipitation estimation using L-band and C-band soil moisture retrievals. *Water Resources Research*, *52*, 7213-7225.

Koster, R. D., Reichle, R. H., & Mahanama, S. P. (2017). A data-driven approach for daily real-time estimates and forecasts of near-surface soil moisture. *J. Hydrometeorology*, *18*, 837-843.

Koster, R. D., Crow, W. T., Reichle, R. H., & Mahanama, S. P. (2018). Estimating basin-scale water budgets with SMAP soil moisture data. *Water Resources Research*, *54*. https://doi.org/10.1029/2018WR022669

Koster, R. D., Reichle, R. H., Schubert, S. D., & Mahanama, S. P. (2019). Length scales of hydrological variability as inferred from SMAP soil moisture retrievals. *J. Hydromet.*, *20*, 2129-2146. doi: 10.1175/JHM-D-19-0070.1

Kummerow, C.D., Randel, D. L., Kulie, M., Wang, N., Ferraro, R., Joseph Munchak, S., & Petkovic, V. (2015). The Evolution of the Goddard Profiling Algorithm to a Fully Parametric Scheme. *J. Atmos. Oceanic Technol*., *32*, 2265–2280. doi:10.1175/JTECH-D-15-0039.1.

Larson, L. W., & Peck, E. L. (1974). Accuracy of precipitation measurements for hydrologic modeling. *Water Resour. Res*., *10*, 857-863.

Lucchesi, R. (2018). File specification for GEOS-5 FP (Forward Processing). GMAO Office Note 4, version 1.2, 56 pp., https://gmao.gsfc.nasa.gov/pubs/docs/Lucchesi1203.pdf.

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

Obled, C., Wendling, J., & Beven, K. (1994). The sensitivity of hydrological models to spatial rainfall patterns, An evaluation using observed data. *J. Hydrol.*, *159*, 305-333. doi: 10.1016/0022-1694(94)90263-1.

Randel D.L., Kummerow, C.D., & Ringerud, S. (2020). The Goddard Profiling (GPROF) Precipitation Retrieval Algorithm. In: Levizzani V., Kidd C., Kirschbaum D., Kummerow C., Nakamura K., Turk F. (eds) *Satellite Precipitation Measurement. Advances in Global Change Research*, vol. *67*. Springer, Cham., doi: 10.1007/978-3-030-24568-9_8

Reichle, R. H., & Q. Liu (2014). Observation-corrected precipitation estimates in GEOS-5. NASA/TM-2014-104606, Vol. 35, 18 pp. [Available online at http://gmao.gsfc.nasa.gov/pubs.]

Reichle, R. H., Liu, Q., Koster, R. D., Draper, C. S., Mahanama, S. P., & Partyka, G. S. (2017a). Land surface precipitation in MERRA-2. *J. Climate*, *30*, 1643-1664. doi: 10.1175/JCLI-D-16-0570.1.

Reichle, R. H., De Lannoy, G. J. M., Liu, Q., Ardizzone, J. V., Colliander, A., Conaty, A., et al. (2017b). Assessment of the SMAP Level-4 surface and root-zone soil moisture product using in situ measurements. *Journal of Hydrometeorology*, *18(10)*, 2621–2645. https://doi.org/10.1175/JHM-D-17-0063.1

Reichle, R. H., Liu, Q., Koster, R. D., Crow, W. T., De Lannoy, G. J. M., Kimball, J. S., et al. (2019). Version 4 of the SMAP level-4 soil moisture algorithm and data product. *J. Adv. Model. Earth Syst.*, *11*, 3106–3130. https://doi.org/10.1029/2019MS001729.

Reichle, R. H., Liu, Q., Ardizzone, J. V., Crow, W. T., De Lannoy, G. J. M., Dong, J., et al. (2021). The contributions of gauge-based precipitation and SMAP brightness temperature observations to the skill of the SMAP Level-4 soil moisture product. *J. Hydromet.*, *22*, 405-424. doi: 10.1175/JHM-D-20-0217.1.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling, The challenge of identifying input and structural errors. *Water Resour. Res*., *46*, 22 pp., doi: 10.1029/2009WR008328.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., & M. Ziese, M. (2015). GPCC Full Data Monthly Product Version 7.0 at 0.5°, Monthly land-surface precipitation from rain-gauges built on GTS-based and historic data. DOI: 10.5676/DWD_GPCC/FD_M_V7_050

Sorooshian, S., Hsu, K.-L., Gao, X., Gupta, H. V., Imam, B., & Braithwaite, D. (2000). Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Amer. Meteorol. Soc*., *81*, 2035-2046. doi: 10.1175/1520-0477(2000)081<2035:EOPSSE>2.3.CO;2 .

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, *J. Geophys. Res*., *103*, 7755–7766. doi:10.1029/97JC03180.

Su, C.-H., Ryu, D., Crow, W. T., & Western, A. W. (2014). Beyond triple collocation: Applications to soil moisture monitoring. *J. Geophys. Res*., *119*, 6419–6439, https://doi.org/10.1002/2013JD021043.

Tapiador, F.J., Turk, F. J., Petersen, W., Hou, A. Y., Garcıa-Ortega, E., Machado, et al. (2012). Global Precipitation Measurement: Methods, Datasets and Applications. *Atmos. Res.*, *104*, 70-97, doi:10.1016/j.atmosres.2011.10.021

Tarpanelli, A., Massari, C., Ciabatta, L., Filippucci, P., Amarnath, G., & L. Brocca, L. (2017). Exploiting a constellation of satellite soil moisture sensors for accurate rainfall estimation. *Advances in Water Resources*, 108, 249-255. doi: 10.1016/j.advwatres.2017.08.010.

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., et al. (2013). The ASCAT soil moisture product: A review of its specifications, validation results, and emerging applications. *Meteor. Z.*, *22*, 5–33. doi:10.1127/0941-2948/2013/0399.

Xie, P., Chen, M., Yang, S., Yatagai, A., Hayasaka, T., Fukushima, Y., & Liu, C. (2007). A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeor*., *8*, 607–626, doi:10.1175/JHM583.1

Figure Captions

Figure 1. Optimal weights to apply to three time series (X_1, X_2, X_3) in producing a merged dataset, as a function of the correlation between each time series and the unknown truth. A full set of contours is shown for three selected values of $\rho_{3,Truth}$: (a) 0.25, (b) 0.5, and (c) 0.75.

Figure 2. a. Triple collocation-based estimates of the square of the temporal correlation between the Gauge/Analysis pentad precipitation data and the unknown truth. White areas indicate where triple collocation-based estimates of accuracy were not possible given data availability (at least 100 samples from all contributors from which to compute correlations). b. Number of gauges per 0.5°x0.5° grid cell in the raw CPCU gauge-based precipitation dataset during the studied period. (Data are plotted here on the 36-km EASE grid; values can be non-integers due to both the combining, through conservative regridding, of different grid cell density numbers into a single grid cell value and to the fact that the values shown represent time averages.) Gauge density in Africa and north of 62.5N is not shown, as the Gauge/Analysis dataset does not utilize rain gauges in these areas (see text). The horizontal lines at 42.5N and 62.5N delimit the area over which the tapered merging of gauge data and analysis data is performed (see Reichle et al. 2017a).

Figure 3. a. Triple collocation-based estimates of the square of the temporal correlation between the IMERG pentad precipitation data and the unknown truth. White areas indicate where triple collocation-based estimates of accuracy were not possible. b. Same, but for the SM2RAIN-based pentad precipitation data.

Figure 4. Triple collocation-based estimates of the maximum skill attainable from the merged precipitation dataset, expressed as the square of the temporal correlation between the merged time series and the unknown truth. White areas indicate where triple collocation-based estimates of accuracy were not possible.

Figure 5. Degree to which the merged precipitation dataset can improve over each of the individual contributors, expressed as the difference between the maximum accuracy (square of the temporal correlation coefficient) for the Merged data shown in Figure 4 minus the accuracy estimates provided for each contributor in Figures 2 and 3. White areas indicate where triple collocation-based estimates of this improvement were not possible. a. Potential improvement of the Merged dataset over the Gauge/Analysis dataset. b. Potential improvement of the Merged dataset over the IMERG dataset. c. Potential improvement of the Merged dataset over the SM2RAIN-based dataset.

Figure 6. Weights applied in the merging process to the (a) Gauge/Analysis dataset, (b) the IMERG dataset, and (c) the SM2RAIN-based dataset. The white dots (two in North America and one in east Asia) indicate locations where sample precipitation time series will be displayed in Figure 7.

Figure 7. Sample time series of pentad precipitation rates for grid cells in: (a) the western US, (b) the upper Midwest US, and (c) eastern Russia. See Figure 6 for specific locations.

Figure 8. Degree to which the Merged dataset improves over each of the contributors when soil moistures generated with each dataset are compared to independent ASCAT measurements. (Skill is measured in terms of anomaly correlations; see Section 2.3.3.) Negative correlations are zeroed prior to squaring. White areas indicate areas for which

comparisons were not possible due to limitations in the triple collocation analysis or to ASCAT data deficiencies. (a) Improvement over the Gauge/Analysis data. (b) Improvement over the IMERG data. (c) Improvement over the SM2RAIN-based data.

Figure 9. Degree to which the Merged dataset improves over each of the contributors when each dataset is correlated against gridded near-surface air temperature (T2M) differences (daily maximum T2M minus daily minimum T2M). Negative correlations are expected, so positive correlations are zeroed prior to squaring. White areas indicate areas for which comparisons were not possible due to limitations in the triple collocation analysis. (a) Improvement over the Gauge/Analysis data. (b) Improvement over the IMERG data. (c) Improvement over the SM2RAIN-based data.

Figure 10. Impact of sample size on the effectiveness of the merging procedure, as revealed by an idealized Monte Carlo analysis (see text). Shown are the average correlations against truth of the merged data as functions of $\rho_{1,Truth}$ and $\rho_{2,Truth}$ (i.e., the prescribed correlations between time series X_1 and X_2, respectively, and the unknown truth); $\rho_{3,Truth}$ for the third time series, X_3, is set to 0.5 for all plots. Top left: results for the case when the weights for the merging are determined from time series with length 120. Top right: Same, but for weights based on time series of length 500. Bottom left: Same, but for weights based on time series of length 1000. Bottom right: Same, but for weights based on (effectively) infinite time series length. The example indicated by the small white dot is discussed in the text.

Figure 11. As in Figure 8, but showing the degree to which the Merged dataset improves over a "consensus" merging of the Gauge/Analysis, IMERG, and SM2RAIN-based data in which each contributor is assigned equal weight (0.33). White areas indicate areas for

which comparisons were not possible due to limitations in the triple collocation analysis or to ASCAT data deficiencies.