https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

Estimating global precipitation fields from rain gauge observations using local ensemble data assimilation

Yuka Muto[1] and Shunji Kotsuki[1, 2, 3, 4]

[1]Center for Environmental Remote Sensing, Chiba University, Chiba, Japan
[2]Institute for Advanced Academic Research, Chiba University, Chiba, Japan
[3]Research Institute of Disaster Medicine, Chiba University, Chiba, Japan
[4]Data Assimilation Research Team, RIKEN Center for Computational Science, Kobe, Japan

Correspondence to: Yuka Muto (yukamuto@chiba-u.jp) and Shunji Kotsuki (shunji.kotsuki@chiba-u.jp)

Abstract. It is crucial to improve global precipitation estimates for a better understanding on water-related disasters and water resources. This study proposes a new methodology to interpolate global precipitation fields from ground rain gauge observations using the algorithm of the local ensemble transform Kalman filter (LETKF) in which the first guess and its error covariance are developed based on the reanalysis data of precipitation from the European Center for Medium-Range Forecasts (ERA5). For the estimation of each date, the climatological ensembles are constructed using the ERA5 data 10 years before and after that date, and thereafter are utilized to obtain the first guess and its error covariance. Additionally, the global rain gauge observations provided by the National Oceanic and Atmospheric Administration Climate Prediction Center (NOAA CPC) are used for observation inputs in the LETKF algorithm.

Our estimates have better agreements against independent rain gauge observations compared to the existing precipitation estimates of the NOAA CPC in general. Because we utilized the same rain gauge observations for the inputs of our estimation as those used in the NOAA CPC product, it is indicated that the proposed estimation method is superior to that of the NOAA CPC (i.e., the Optimal Interpolation). Our proposed method took the advantage of constructing a physically guaranteed first guess and its error variance using reanalysis data for interpolating precipitation fields. Furthermore, the method of this study is shown to be particularly beneficial for mountainous or rain-gauge-sparse regions.

1 Introduction

Improving the accuracy of global precipitation fields is crucial for predicting water-related disasters such as floods and droughts, and for long-term water resource management. Ground rain gauge observations play an essential role in estimating global precipitation fields, because it is considered to be more accurate relative to other estimates by numerical weather prediction (NWP) models or satellite-borne sensors, especially in mountainous areas (Sun et al. 2018). On the other hand, rain gauge observations can only be acquired at a limited number of locations. The National Oceanic and Atmospheric Administration Climate Prediction Center (NOAA CPC) provides the CPC Unified Gauge-based Analysis of Global Daily Precipitation (hereafter, CPC_est) (Xie et al. 2007; Chen et al., 2008), which is spatially interpolated precipitation data based

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

on rain gauge observations. Such global precipitation data are important not only as input data to analyze the hydrological water cycle, but as a reference data for validating or adjusting NWPs and satellite-based precipitation estimates. For example, the satellite-based Global Satellite Mapping of Precipitation (GSMaP), which is provided by the Japan Aerospace Exploration Agency (Kubota et al., 2020), is adjusted to CPC_est (Mega et al., 2019). Thus, even with the advancements in satellite observations and numerical weather forecasting, the methodology to improve global precipitation fields by utilizing precise ground rain gauge observations is demanding.

There have been many methodological studies to estimate precipitation fields from sparsely located rain gauge observations (e.g., Cressman, 1959; Barnes, 1964; Gandin, 1965; Shepard, 1968). Among them, a widely used interpolation method is the Optimal Interpolation (OI) (Gandin, 1965), which provides a weighted average of the first guess on each grid point and the surrounding observations. Because the OI determines the weights of the first guess and observations by considering the error variance and covariance as well as the distance with respect to the surrounding observation points, this method was suggested to be superior to the other inverse-distance weighting methods of Cressman (1959) and Shepard (1968) (Chen et al, 2002). Consequently, the operational global precipitation fields of CPC_est uses the OI to the present day (Xie et al. 2007).

In recent years, more sophisticated interpolation methods have been introduced from the field of data assimilation. For example, Kumar et al. (2021) applied a data assimilation approach to combine the satellite-based GSMaP and rain gauge observations in India, using GSMaP and rain gauge observations as the first guess and the observation inputs, respectively. The proposed method in Kumar et al. (2021) constructs a flow-dependent background error covariance by implementing the Kalman filter (Kalman, 1960) to propagate the background error covariance. Furthermore, the accuracy of NWPs has improved rapidly over the past few decades (Pu and Kalnay, 2018). Because NWP-based data capture dynamical relationships between locations and variables, rain-gauge-based precipitation estimates would be further improved by using NWP-based data for the first guess and background error covariance. Here, ensemble data assimilation (EnDA) can be used to obtain the climatological error covariance by regarding NWP-based precipitation records as an ensemble (Kretschmer et al. 2015; Kotsuki and Bishop 2022). In particular, the Local Ensemble Transform Kalman Filter (LETKF) (Hunt et al., 2007) is a computationally efficient EnDA method which extracts the observations close to the grid point by a localization method, and has been implemented in many previous studies on NWPs (e.g., Hamrud et al., 2015; Terasaki et al., 2015; Schraff et al., 2016). Hence, this study aims to propose a new estimation method for global precipitation fields by spatial interpolation from rain gauge observations, utilizing the LETKF algorithm and NWP-based data. Furthermore, we will verify the superiority of our estimation method with comparison to the OI used in CPC_est.

The rest of the paper is organized as follows. Section 2 describes the proposed interpolation method, followed by the validation methods with respect to independent rain gauge observation data. Section 3 presents the precipitation fields estimated by the proposed method as well as the results of the validations. The advantage of the proposed method are discussed in Section 4, followed by a conclusion in Section 5.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

2 Methods

2.1 Interpolation method

This section describes the interpolation method whose schematic image is shown in Fig. 1.

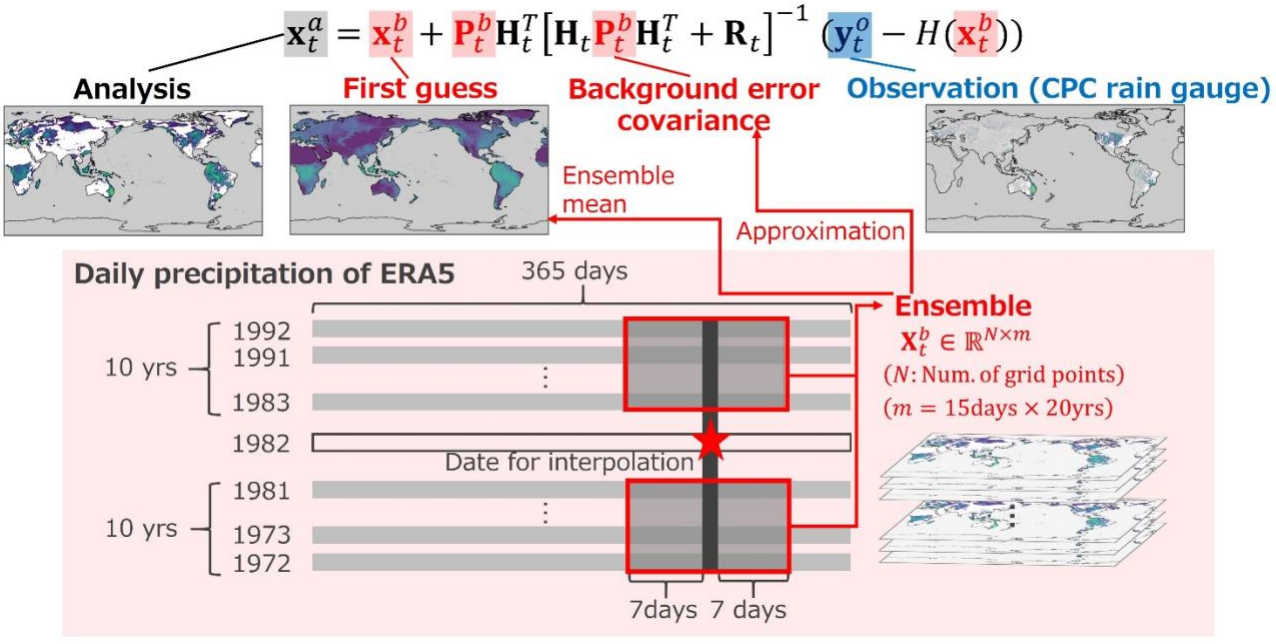

Figure 1: The schematic image of the interpolation method of this study using ensemble data assimilation. The rain gauge observations from the CPC product are used for the observation $\mathbf{y}_t^o$. The ensemble $\mathbf{X}_t^b$ is obtained from the daily precipitation data from the fifth generation ECMWF reanalysis (ERA5) before and after the interpolation date, and the ensemble mean is used as the first guess $\mathbf{x}_t^b$. $\mathbf{R}_t$ is the observation error covariance. H_t denotes an observation operator that maps the first guess values to the observed values, and $\mathbf{H}_t$ is the Jacobi matrix of $H_t(\mathbf{x}_t^b)$. The background error covariance $\mathbf{P}_t^b$ is also approximated from the ensemble. Finally, the interpolated daily global precipitation field is computed as the analysis $\mathbf{x}_t^a$.

2.1.1 Input Data

This study uses the rain gauge data utilized in CPC_est for the observation input for the interpolations. These rain gauge data are collected by NOAA CPC from approximately 16,000 stations, including daily summary files from the Global Telecommunication System (GTS) and the CPC unified daily precipitation data sets over the contiguous United States, Mexico and South America (Chen et al., 2008). Since CPC_est is published as a 0.5° × 0.5° pixel data, we only use the precipitation at pixels in which more than one rain gauge station is included (hereafter, CPC_gauge), and also assume that the rain gauge(s) is (are) located at the center of each pixel.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

For the construction of the first guess and background error covariance, we use the "Total precipitation" data from the fifth generation ECMWF reanalysis (ERA5) (Hersbach et al., 2023). ERA5 is a 0.25° × 0.25° gridded hourly data based on the Integrated Forecast System (version Cy41r2), and combined with various conventional and satellite observations related to atmosphere, land and ocean by data assimilation (Hersbach et al., 2020). We computed the total precipitation on a daily basis from the original ERA5 data. Although the original ERA5 data cover both land and sea areas, this study focused on estimating the precipitation fields only over land, where rain gauge observations are available.

2.1.2 Ensemble data assimilation

The daily precipitation in the same grid points as ERA5 over land is estimated using CPC_gauge as observation inputs according to Equation (1), which is the equation of the Kalman filter (Kalman, 1960):

$$\mathbf{x}_t^a = \mathbf{x}_t^b + \mathbf{P}_t^b \mathbf{H}_t^{\mathrm{T}} [\mathbf{H}_t \mathbf{P}_t^b \mathbf{H}_t^{\mathrm{T}} + \mathbf{R}_t]^{-1} \left(\mathbf{y}_t^o - H_t(\mathbf{x}_t^b) \right), \tag{1}$$

where $\mathbf{x}_t^a \in \mathbb{R}^N$, $\mathbf{x}_t^b \in \mathbb{R}^N$, $\mathbf{y}_t^o \in \mathbb{R}^P$ denote the analysis, first guess, and observation values at time t, respectively. Superscripts a, b and o denote the analysis, first guess, and observation, respectively. $\mathbf{P}_t^b \in \mathbb{R}^{N \times N}$ and $\mathbf{R}_t \in \mathbb{R}^{P \times P}$ represent the background and observation error covariance. The scalers N and P denote the number of grid points of ERA5 over land, and that of CPC_gauge pixels, respectively. H_t denotes an observation operator that maps the first guess to the observed values, and $\mathbf{H}_t \in \mathbb{R}^{P \times N}$ is the Jacobi matrix of $H_t(\mathbf{x}_t^b)$.

Here, we define $\mathbf{R}_t$ as a diagonal matrix owing to the assumption that the errors of the observations are independent from each other. The error variance of each observation (i.e., the diagonal components of $\mathbf{R}_t$) is given by Equation (2), which is determined based on preliminary sensitivity experiments:

$$\text{error variance} = \begin{cases} \log(2) & \left(y_{l,t}^o \le 1.0 \text{ mm day}^{-1}\right) \\ \log\left(y_{l,t}^o + 1\right) & \left(y_{l,t}^o > 1.0 \text{ mm day}^{-1}\right) \end{cases}, \tag{2}$$

where log is the natural logarithm and $y_{l,t}^o$ denotes the observation at the lth pixel and tth time step from CPC_gauge.

The first guess values of $\mathbf{x}_t^b$ and the background error covariance $\mathbf{P}_t^b$ are given by the daily precipitation of ERA5. For each estimation date, the data of the 10 years before and after that date is extracted. Then, we extract the data of the same day of year as the estimation date and also the surrounding 7 days within those 20 years, and utilize them as an ensemble $\mathbf{X}_t^b$ (cf. Fig. 1). We do not extract the ERA5 data in the exact year of the estimation date, because we compare our precipitation estimates with ERA5 itself for validation (details are explained in Section 2.2.2). Thereafter, the first guess $\bar{\mathbf{x}}_t^b$ is given by the mean of the ensemble. Additionally, $\mathbf{P}_t^b$ is approximated by the ensemble (Evensen,1994), given by:

$$\mathbf{P}_t^b \approx \mathbf{Z}_t^b (\mathbf{Z}_t^b)^{\mathrm{T}}, \tag{3}$$

$$\mathbf{Z}_t^b = \frac{\delta \mathbf{X}_t^b}{\sqrt{M-1}}, \tag{4}$$

where $\delta \mathbf{X}_t^b \in \mathbb{R}^{N \times M}$ denotes the ensemble perturbation between the respective ensemble and the ensemble mean for each grid, and M denotes the number of ensembles ($M = 15 \text{ days} \times 20 \text{ yrs}$).

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

Ensemble data assimilation usually requires the localization so that the observation values are weighted according to their distance from the analysis grid point using the localization function. When the distance between a grid point in the first guess and an observation site is d km, the localization function $L(d)$ is expressed by Equation (5):

$$L(d) = \begin{cases} \exp\left(-\frac{d^2}{2\sigma^2}\right) & d < 2\sqrt{10/3}\,\sigma \\ 0 & \text{else} \end{cases}, \tag{5}$$

where σ denotes the localization scale (km). Localization is performed by dividing the diagonal component of $\mathbf{R}_t$ by $L(d)$ for each grid point and observation site, so that observations distant from the analysis grid point have less weights. Here, we determine the value of σ based on the method of Schraff et al. (2016), known as the Observation Number Limit technique. First, a certain distance d_{max}^{ini} km is set, followed by the maximum number of observation sites (P_{loc}^{max}) to be used for the estimation. Next, the localization scale σ is determined by Equation (6):

$$\sigma = \begin{cases} \frac{d_{max}^{ini}}{2\sqrt{10/3}} & P_{loc}^{ini} < P_{loc}^{max} \\ \frac{d_{max}^{fix}}{2\sqrt{10/3}} & \text{else} \end{cases}, \tag{6}$$

where P_{loc}^{ini} denotes the number of observation sites within the d_{max}^{ini} km radius from the grid point, and d_{max}^{fix} is the distance (in km) between the grid point and the $(P_{loc}^{max}+1)$th nearest observation site. The tunable parameters d_{max}^{ini} and P_{loc}^{max} are set to 1,000 km and 10 respectively, owing to the authors' preliminary experiments.

Our study applies the LETKF algorithm, in which the ensemble mean of the analyses $\bar{\mathbf{x}}_t^a$ is computed by Equation (7) (Hunt et al, 2007):

$$\bar{\mathbf{x}}_t^a = \bar{\mathbf{x}}_t^b + \mathbf{Z}_t^b \tilde{\mathbf{P}}_t^a (\mathbf{H}_t \mathbf{Z}_t^b)^{\mathrm{T}} \mathbf{R}_{t\,loc}^{-1} \left(\mathbf{y}_t^o - H_t(\mathbf{x}_t^b)\right), \tag{7}$$

where $\mathbf{R}_{t\,loc}^{-1} \in \mathbb{R}^{P_{loc} \times P_{loc}}$ denote the inverse of $\mathbf{R}_t$ with the localization, and $\mathbf{I}$ denote the identity matrix. The scaler P_{loc} denotes the number of observations within the localization cut-off radius. Here, we compute $\tilde{\mathbf{P}}_t^a$ using the following equations proposed by Kotsuki and Bishop (2022):

$$\tilde{\mathbf{P}}_t^a = \mathbf{C}(\mathbf{I} + \mathbf{\Gamma})^{-1}\mathbf{C}^{\mathrm{T}}, \tag{8}$$

$$\mathbf{C} = (\mathbf{H}_t \mathbf{Z}_t^b)^{\mathrm{T}} \mathbf{R}_{t\,loc}^{-1/2} \mathbf{E} \mathbf{\Gamma}^{-1/2}, \tag{9}$$

where the eigenvalue decomposition is solved for a $P_{loc} \times P_{loc}$ matrix given by:

$$\mathbf{R}_{t\,loc}^{-1/2} \mathbf{H}_t \mathbf{Z}_t^b (\mathbf{H}_t \mathbf{Z}_t^b)^{\mathrm{T}} \mathbf{R}_{t\,loc}^{-1/2} = \mathbf{E}\mathbf{\Gamma}\mathbf{E}^{\mathrm{T}}. \tag{10}$$

Because the number of local observations $P_{loc}(\leq 10)$ is smaller than the ensemble size $M\,(=300)$, the computational cost is smaller than the original LETKF algorithm, in which the eigenvalue decomposition is solved for an $M \times M$ matrix $\left(\tilde{\mathbf{P}}_t^a\right)^{-1}$.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

Consequently, $\bar{\mathbf{x}}_t^a$ is the interpolated daily global precipitation field, and is used as the final estimate of this study (hereafter, LETKF_est). Based on the method explained above, we estimated the daily global precipitation field for ten years (1981−1990). Note that we skip the estimation for 23 days during the estimation period when no valid rain gauge observations are available in either Africa, Eurasia or Canada.

2.2 Validations

2.2.1 Data used for the validations

Two precipitation data are used for the validations. The first data is APHRODITE (Yatagai et al., 2012), which is also a daily precipitation dataset constructed by applying interpolation based on rain gauge observations. In addition to the rain gauge data from the GTS, APHRODITE uses rain gauge data precompiled by other projects or organizations and those originally collected from national hydrological and meteorological services, therefore enabling validations against rain gauge observations independent from those used in CPC_est. Here, we use the 0.5˚ × 0.5˚ pixel data of the latest version of APHRODITE (V1101) in Monsoon Asia (MA) (Fig. 2 a), where particularly dense rain gauge data are available compared to those from the GTS.

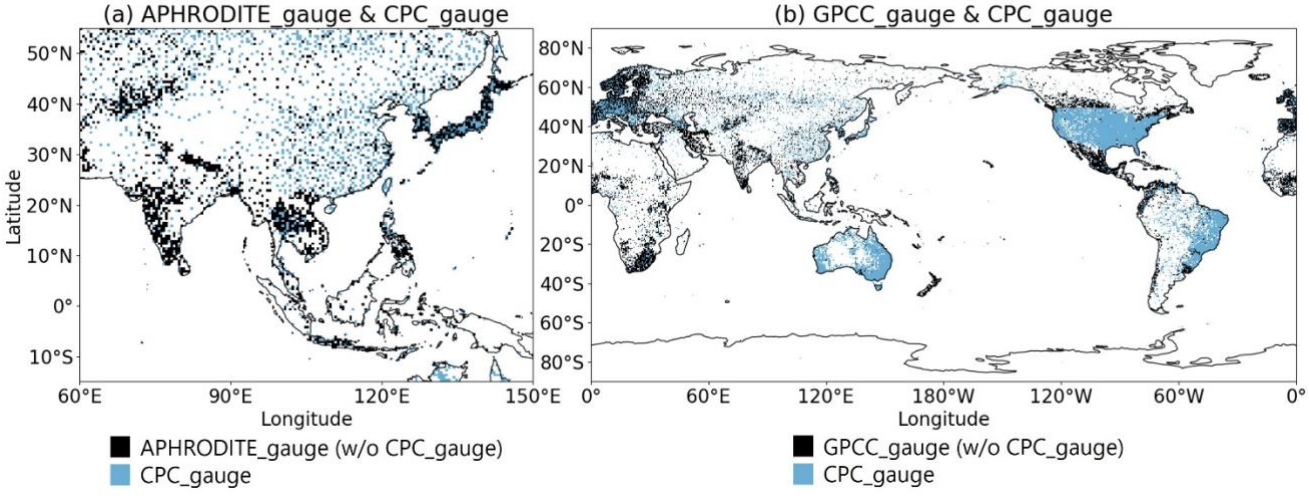

Figure 2: An example of (a) the distribution of the daily rain gauge observations used in APHRODITE v1101 and the CPC product in Monsoon Asia (on Nov. 15th, 1988), and (b) the monthly rain gauge observations used in the GPCC FD product v2022 and the CPC product (in Nov., 1988). The black pixels include more than one rain gauge stations which are independent from the stations used in the CPC product, and the light blue pixels include more than one rain gauge stations used in the CPC product.

Secondly, the monthly precipitation product of the Global Precipitation Climatology Centre (GPCC) is used. The Full Data Reanalysis (FD) product of the GPCC is constructed based on rain gauge observations from > 40,000 stations throughout the globe, including not only the observations used in CPC_est, but also data provided from other sources such as the national

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

data by the World Meteorological Organization or the collection of the Global Historical Climatology Network (Becker et al., 2013). Thus, although in a monthly basis, the GPCC is used as rain gauge observations independent from CPC_gauge in a global scale (Fig. 2 b). In this study, we use the latest version of the 0.5˚ × 0.5˚ pixel FD product (v2022) (Schneider et al., 2022).

For both the APHRODITE and GPCC products, we use the data samples of the pixels in which more than one rain gauge is included (APHRODITE_gauge and GPCC_gauge), and assume that the rain gauge(s) is (are) located at the center of each pixel, similar to CPC_gauge. Prior to the validations, the 0.25˚ × 0.25˚ gridded LETKF_est and ERA5 data are converted into 0.5˚ × 0.5˚ pixel data so as to be equivalent to the spatial resolution of CPC_est, APHRODITE_gauge and GPCC_gauge.

2.2.2 Validation against APHRODITE_gauge

Here, we use an index that measures correlation based on the rank of the samples rather than the exact magnitude of them, considering that some studies have suggested the possibility of the APHRODITE precipitation to be biased (Kotsuki and Tanaka, 2013; Ji et al., 2020). Such index is also less susceptible to low-frequency extreme values, which may occur in daily precipitation data. Hence, Kendall's rank correlation coefficient τ_b (Kendall, 1948) is computed against the daily precipitation of APHRODITE_gauge for LETKF_est, CPC_est, and ERA5, respectively. When N_{aphro} is the number of APHRODITE_gauge pixels, and (u_i, v_i) $(i = 1, \dots, N_{aphro})$ are the pairs of data to be compared, τ_b is obtained by Equation (11) and (12):

$$\tau_b = \frac{A - B}{\sqrt{S - T_u}\sqrt{S - T_v}}, \tag{11}$$

$$S = \frac{N_{aphro}\left(N_{aphro} - 1\right)}{2}, \tag{12}$$

where A (B) represent the total number of cases in which the magnitude correlation of u_j $(j = 1, \dots, N_{aphro})$ and u_k $(k = j + 1, \dots, N_{aphro})$ is concordant (discordant) with that of v_j and v_k. T_u and T_v denote the number of ties in u_i and v_i, respectively.

The value of τ_b closer to 1.0 (–1.0) indicates stronger positive (negative) correlation between the two types of data. We exclude the samples of the pixels where the input observations from CPC_gauge are available to evaluate only the interpolated precipitation in our study. Furthermore, we exclude the samples of the pixels where the precipitation of APHRODITE_gauge is < 0.5 mm day^{-1}, considering that precipitation under this value generally cannot be measured precisely by rain gauges.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

2.2.3 Validations against GPCC_gauge

The spatial root mean square difference (RMSD), mean absolute difference (MAD) and Pearson's correlation coefficient (R) are computed for each month against the monthly precipitation of GPCC_gauge for LETKF_est and CPC_est following Equations (13) to (15):

$$Spatial\ RMSD_t = \sqrt{\frac{\sum_{i=1}^{N_{gpcc}} w_i \left(x_{ref\ i,t} - x_{est\ i,t}\right)^2}{\sum_{i=1}^{N_{gpcc}} w_i}}, \tag{13}$$

$$Spatial\ MAD_t = \frac{\sum_{i=1}^{N_{gpcc}} w_i \left|x_{gpcc\ i,t} - x_{est\ i,t}\right|}{\sum_{i=1}^{N_{gpcc}} w_i}, \tag{14}$$

$$R_t = \frac{\frac{1}{N_{gpcc}} \sum_{i=1}^{N_{gpcc}} \left(x_{gpcc\ i,t} - \overline{x_{gpcc\ t}}\right)\left(x_{est\ i,t} - \overline{x_{est\ t}}\right)}{\sqrt{\frac{1}{N_{gpcc}} \sum_{i=1}^{N_{gpcc}} \left(x_{gpcc\ i,t} - \overline{x_{gpcc\ t}}\right)^2} \sqrt{\frac{1}{N_{gpcc}} \sum_{i=1}^{N_{gpcc}} \left(x_{est\ i,t} - \overline{x_{est\ t}}\right)^2}}, \tag{15}$$

where N_{gpcc} denote the number of GPCC_gauge pixels. $x_{gpcc\ i,t}$ and $x_{est\ i,t}$ denote the monthly precipitation of GPCC_gauge and the estimates (LETKF_est or CPC_est) at the ith pixel and tth time step, respectively. Additionally, $\overline{x_{gpcc\ t}}$ and $\overline{x_{est\ t}}$ denote the spatial mean monthly precipitation of GPCC_gauge and the estimates (LETKF_est or CPC_est) at the tth time step, respectively. Here, $w_i = \cos(\theta_i)$ is the latitude-dependent weight of the ith pixel, where θ is the latitude.

Smaller RMSD or MAD values (at the minimum of 0.0) indicate that the two data are similar, while the R value closer to 1.0 (–1.0) indicates stronger positive (negative) correlation. As explained in Section 2.2.2, we also exclude the samples of the pixels where the input observations from CPC_gauge are available for the validations against GPCC_gauge. Additionally, the months in which we skipped the estimation for daily precipitation (as noted in Section 2.1.2) were also excluded from the validations (Jan., 1981; Apr., 1983; Jan., 1984; Jan. –Feb. and Jul. –Aug., 1985.; Jan., Mar., Sep. and Nov., 1986).

3 Results

A first guess precipitation field used in our study, CPC_gauge and LETKF_est on Nov. 15th in1988 is illustrated as an example in Fig. 3 a, b and d, respectively. The daily precipitation field of LETKF_est (Fig. 3 d) is interpolated using the smooth and averaged climatological first guess (Fig. 3 a) and the sparsely located rain gauge observations (Fig. 3 b), using the methodology presented in Section 2.1.2. For the same date, the daily precipitation of NOAA's CPC_est, which is estimated by the OI also using the rain gauge observations in CPC_gauge, is depicted in Fig. 3 c for comparison. Although the precipitation patterns of CPC_est (Fig. 3 c) and LETKF_est (Fig. 3 d) are overall similar to each other, several differences exist between them. For example, broader precipitating areas are seen for LETKF_est than for CPC_est, especially around the central part of Africa,

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

the Himalayas, the Zagros mountains, and the Indochina Peninsula. In addition, the precipitation is generally weaker for LETKF_est than CPC_est.

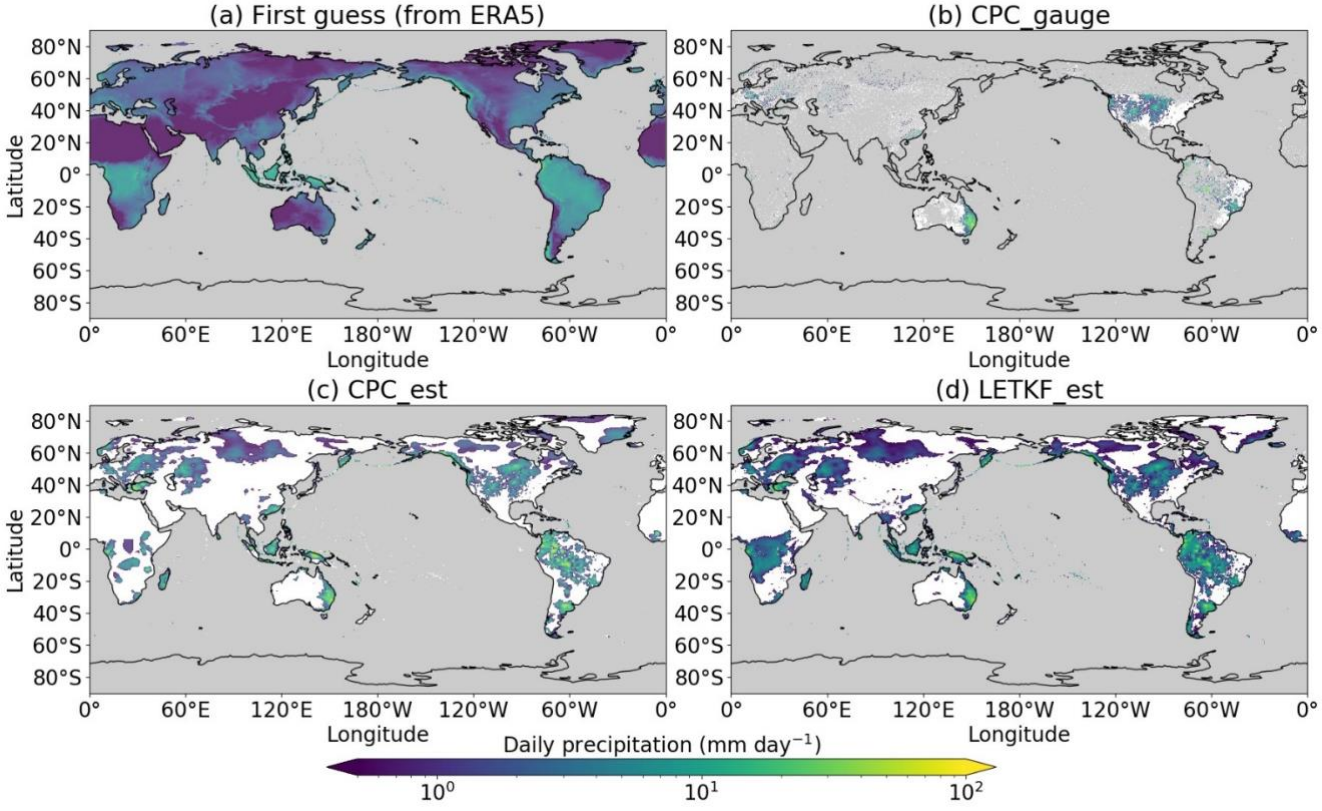

Figure 3: An example of the precipitation fields (mm day^{-1}) for (a) the first guess used in our study, (b) the rain gauge observations of CPC_gauge, and the global precipitation estimates of (c) CPC_est and (d) LETKF_est (on Nov. 15th, 1988). Pixels on the ocean are colored in gray for all subplots, as well as those where no rain gauge observations are available for Subplot (b). Pixels are colored in white when the precipitation is < 0.5 mm day^{-1}.

The scatter plots in Fig. 4 compare the daily precipitation of ERA5, CPC_est and LETKF_est with APHRODITE_gauge at pixels in MA, showing that LETKF_est is aligned with APHRODITE_gauge the most compared to ERA5 and CPC_est. Furthermore, the τ_b value of LETKF_est computed against APHRODITE_gauge is the highest (Fig. 4), notwithstanding that LETKF_est was converted to $0.5^\circ \times 0.5^\circ$ pixel data in advance of this validation. Therefore, it shows that the daily precipitation of LETKF_est is more similar to that of APHRODITE_gauge than ERA5 or CPC_est in terms of Kendall's rank correlation coefficient.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

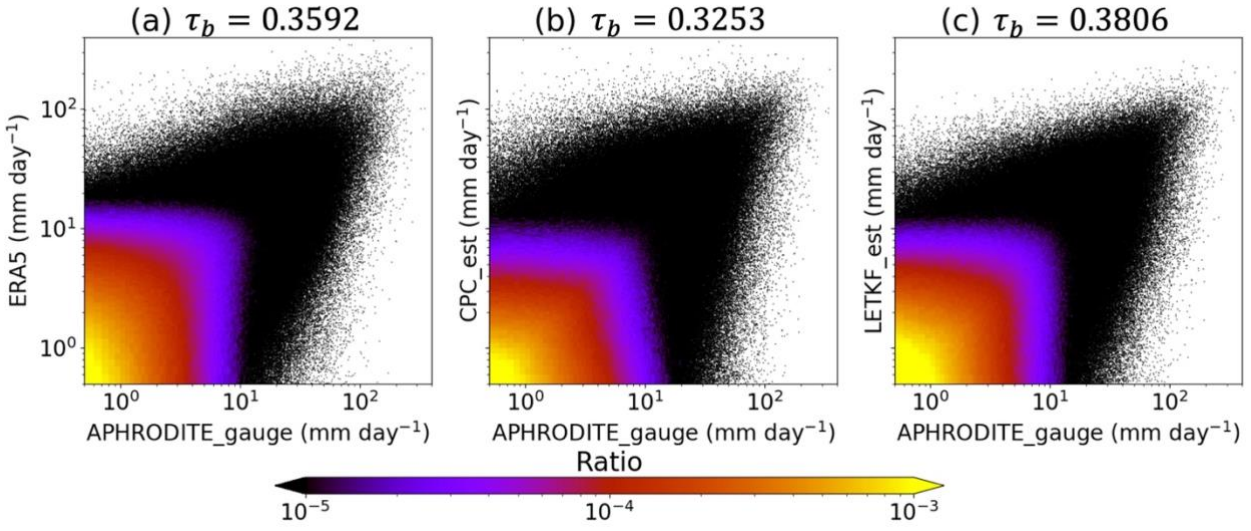

Figure 4: Scatter plots comparing the daily precipitation (mm day^{-1}) of APHRODITE_gauge with that of (a) ERA5, (b) CPC_est and (c) LETKF_est. The colors represent the ratio of samples within each 0.1 mm day^{-1} bin. Kendall's rank correlation coefficient (τ_b) of (a) ERA5, (b) CPC_est and (c) LETKF_est computed against APHRODITE_gauge are listed at the top of each subplot.

The spatial RMSD, MAD and R verified against GPCC_gauge indicate that the monthly precipitation of LETKF_est shows better agreements with GPCC_gauge (i.e., lower RMSD and MAD values, and higher R values) than that of CPC_est for all months throughout the estimation period (Fig. 5). The temporal average of the spatial RMSD and MAD of the LETKF_est is lower than those of CPC_est by 14.79 % and 10.96 %, respectively. The spatial MAD is also computed separately among the low-latitude region (20˚N–20˚S) and mid- and high-latitude regions (90˚N–20˚N and 20˚S–90˚S) against GPCC_gauge for both LETKF_est and CPC_est for each month. Figure 6 indicates that the MAD values in the low-latitude region are generally higher than those of the mid- and high- latitude regions. However, the scatter plots for the low-latitude region are more divergent from the 1:1 line upwards, indicating that the MAD values have improved for LETKF_est compared to CPC_est particularly in this region. Therefore, it is indicated that our estimation method is more beneficial than the OI especially for the low-latitude region, which is highly occupied by the tropical regions with more precipitation.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

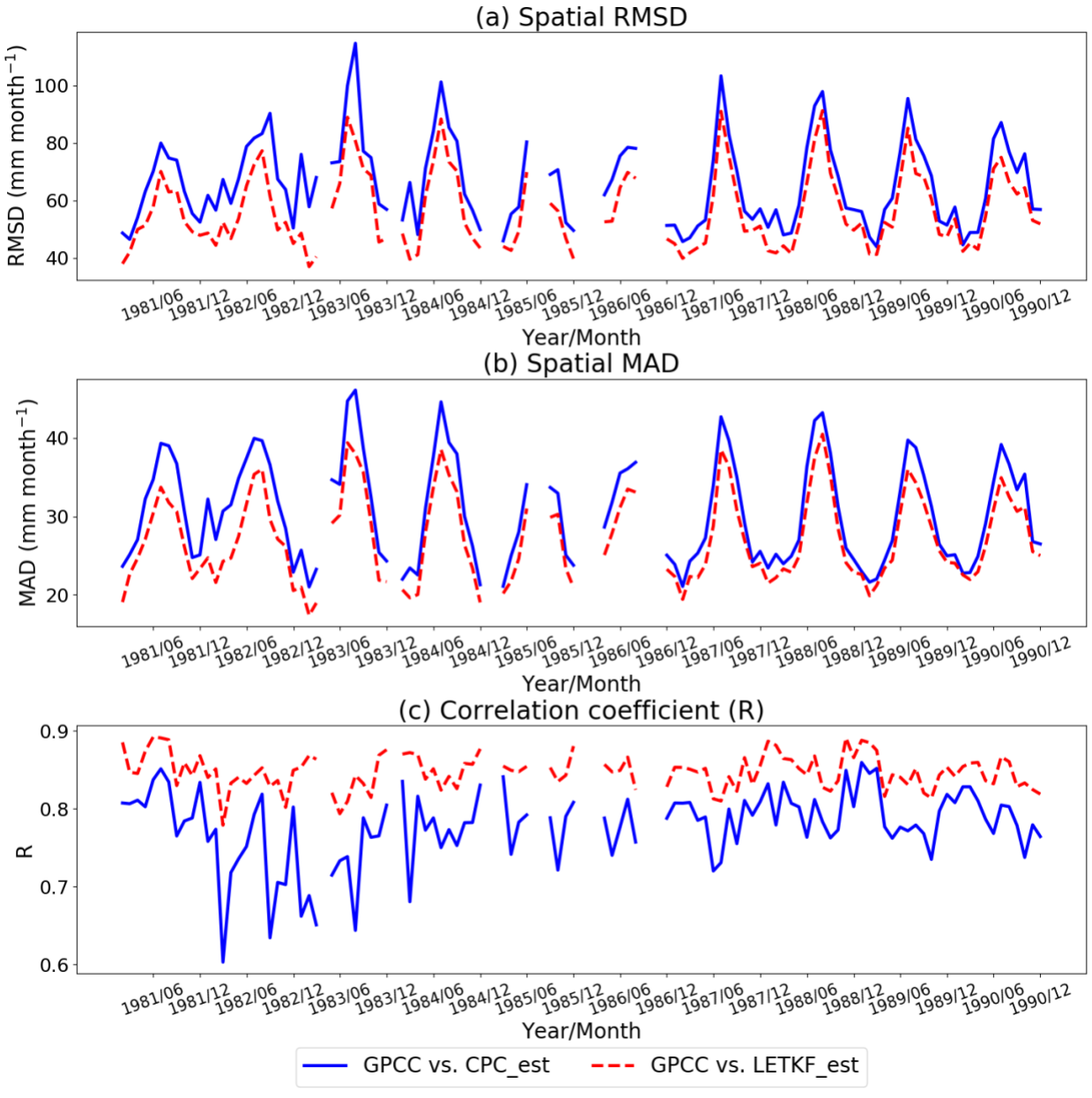

Figure 5: The time series of (a) the spatial root mean square difference (RMSD; mm month^{-1}), (b) the spatial mean absolute difference (MAD; mm month^{-1}) and (c) Pearson's correlation coefficient (R), verified against the GPCC_gauge. The blue solid and red dashed lines represent the CPC_est and LETKF_est, respectively. The validations are not performed for the months in which we skipped the estimation for daily precipitation (Jan., 1981; Apr., 1983; Jan., 1984; Jan. –Feb. and Jul. –Aug., 1985.; Jan., Mar., Sep. and Nov., 1986).

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

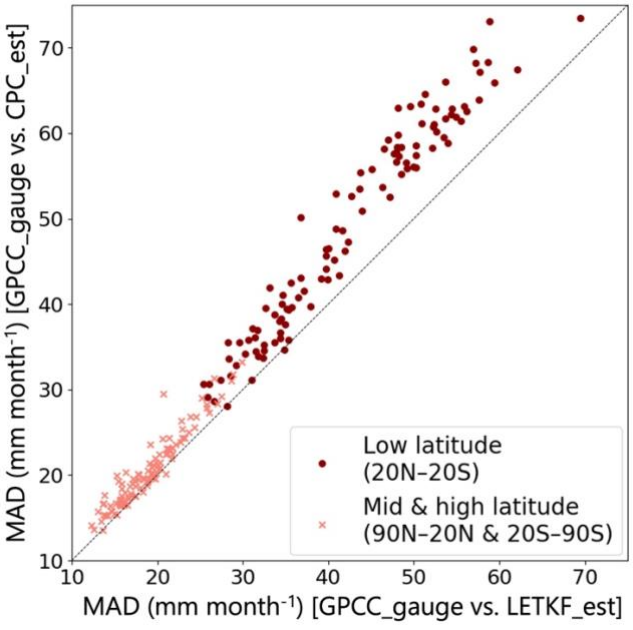

Figure 6: Scatter plots comparing the spatial mean absolute difference (MAD; mm month^{-1}) of CPC_est and LETKF_est verified against the monthly precipitation of GPCC_gauge. Light-red cross marks and dark-red circles represent the low-latitude region (20˚N–20˚S) and mid- and high-latitude regions (90˚N–20˚N and 20˚S–90˚S), respectively.

4 Discussion

The main reason for the improvement in the accuracy of LETKF_est compared to CPC_est is presumably owing to the interpolation method that uses the dynamically guaranteed first guess and background error covariance constructed from the ERA5 data. This would have led to the improvement in the accuracy of the first guess, as well as the variance of each grid point and the covariance between paired grid points. For example, our first guess would take into account the orographic effects. Here, we investigate the difference in Southeast Asian precipitation.

Figure 7 depicts the first guess used for this study and the daily precipitation of CPC_gauge, ERA5, CPC_est and LETKF_est on Jun. 7th, 1985. It should be noted that the precipitation of LETKF_est (Fig. 7 f) is the one converted into a 0.5˚ × 0.5˚ pixel data, for the comparison with that of CPC_est (Fig. 7 e). LETKF_est succeeds in reproducing the orographic changes in precipitation around the Himalayas (Fig. 7 f), while CPC_est fails to do so (Fig. 7 e). The first guess constructed by ERA5 (Fig. 7 b) is presumed to contribute to these precipitation patterns of LETKF_est, since it clearly reflects orographic features, similar to the original ERA5 (Fig. 7 d). On the other hand, as explained in Section 3, the precipitation of LETKF_est

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

265 has better agreement with APHRODITE_gauge than that of ERA5 itself, suggesting that not only the first guess, but also the climatological background error covariance constructed from ERA5 contributes to the improvement in our estimates.

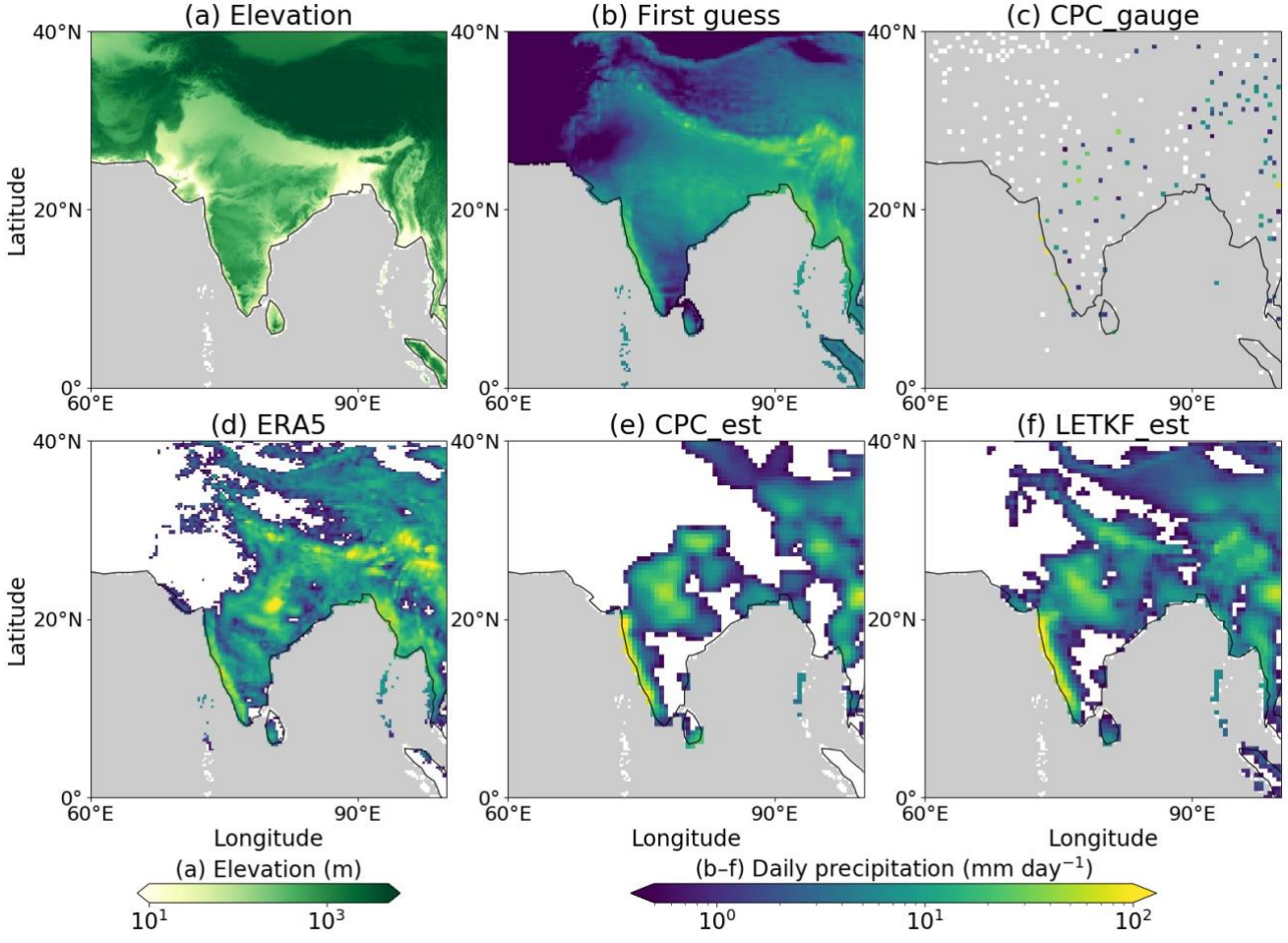

Figure 7: (a) The elevation (m) and an example (b) the first guess constructed in our study (mm day^{-1}), (c) the rain gauge observations of CPC_gauge (mm day^{-1}), and the global precipitation estimates (mm day^{-1}) of (d) ERA5, (e) 270 **CPC_est and (f) LETKF_est (on Jun. 27th, 1985) around India. Pixels on the ocean are colored in gray for all subplots, as well as those where no rain gauge observations are available for Subplot (c). The precipitation of LETKF_est (Subplot (f)) is the one converted into a 0.5° × 0.5° pixel data. Pixels are colored in white when the precipitation is < 0.5 mm day^{-1}.**

275 To investigate whether the precipitation of LETKF_est is more accurate than that of CPC_est around mountainous areas such as the Himalayas in general, Kendall's rank correlation coefficient (τ_b) was computed for LETKF_est and CPC_est against the daily precipitation of APHRODITE_gauge for each pixel where more than 1,800 samples of APHRODITE_gauge

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

are available. The results in Fig. 8 show that the τ_b values of LETKF_est are higher than that of CPC_est by > 0.05 especially around the Himalayas, indicating that the method of this study improves the daily precipitation significantly around this area during the estimation period in general.

Additionally, the temporal MAD values of LETKF_est and CPC_est are computed against the monthly precipitation of GPCC_gauge for each pixel where more than 50 samples of GPCC_gauge are available, using Equation (16):

$$Temporal\ MAD_i = \frac{\sum_{t=1}^{T} \left| x_{ref\ i,t} - x_{est\ i,t} \right|}{T}, \tag{16}$$

where T is the total number of time steps.

The temporal MAD of LETKF_est is smaller than that of CPC_est by > 10 mm month^{-1} at many pixels around mountainous areas such as the Himalayas (Fig. 9 c) and the Zagros Mountains (Fig. 9 f), indicating that the estimation method of this study is beneficial for these areas in general. Furthermore, the temporal MAD of LETKF_est decreased by > 10 mm month^{-1} compared to that of CPC_est in regions where rain gauge stations are especially sparse, such as South-east Asia (Fig. 9 c) or the central part of Africa (Fig. 9 f). In both the mountainous and rain-gauge-sparse regions, the temporal MAD is relatively high compared to other regions (Fig. 9 a–b and d–e). Therefore, although interpolating precipitation fields in such areas is especially difficult, it is presumed that the proposed method succeeded in improving the accuracy of the estimates compared to the conventionally used OI method.

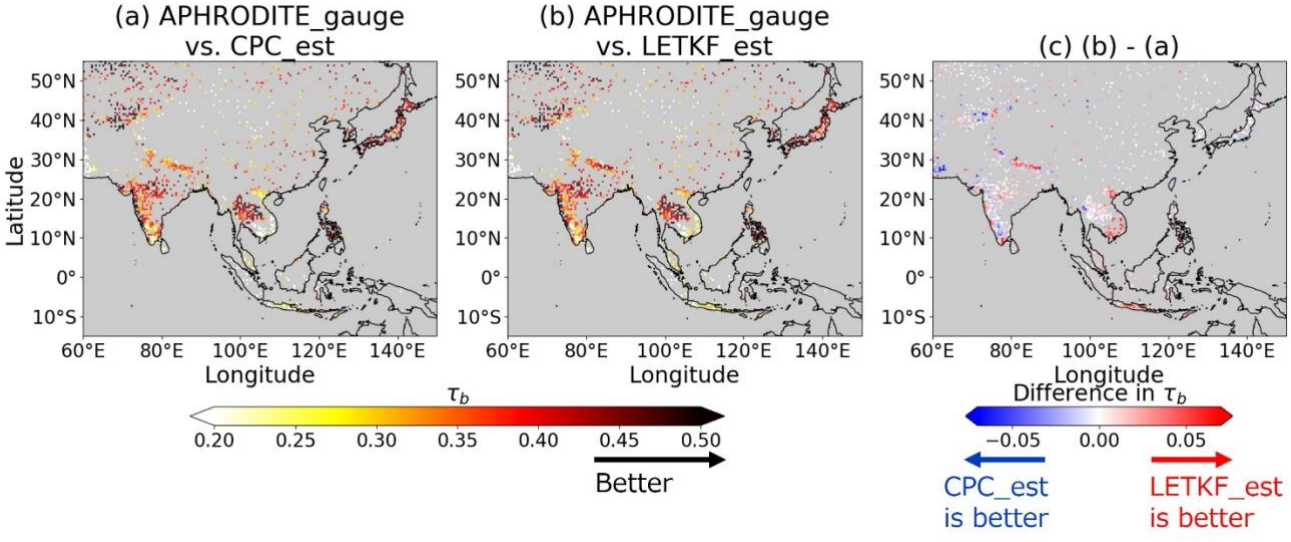

Figure 8: Kendall's rank correlation coefficient (τ_b) computed against the daily precipitation of APHRODITE_gauge for (a) CPC_est and (b) LETKF_est at each pixel. Subplot (c) represents the difference between (b) and (a). Darker colors in (a–b) indicate that the precipitation estimates are more similar to APHRODITE_gauge. Warm colors in (c) indicate that LETKF_est is more similar to APHRODITE_gauge than CPC_est, and cold colors indicate otherwise. τ_b is computed only at pixels where more than 1,800 samples from APHRODITE_gauge are available, and the pixels are colored in gray if they do not match this condition.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

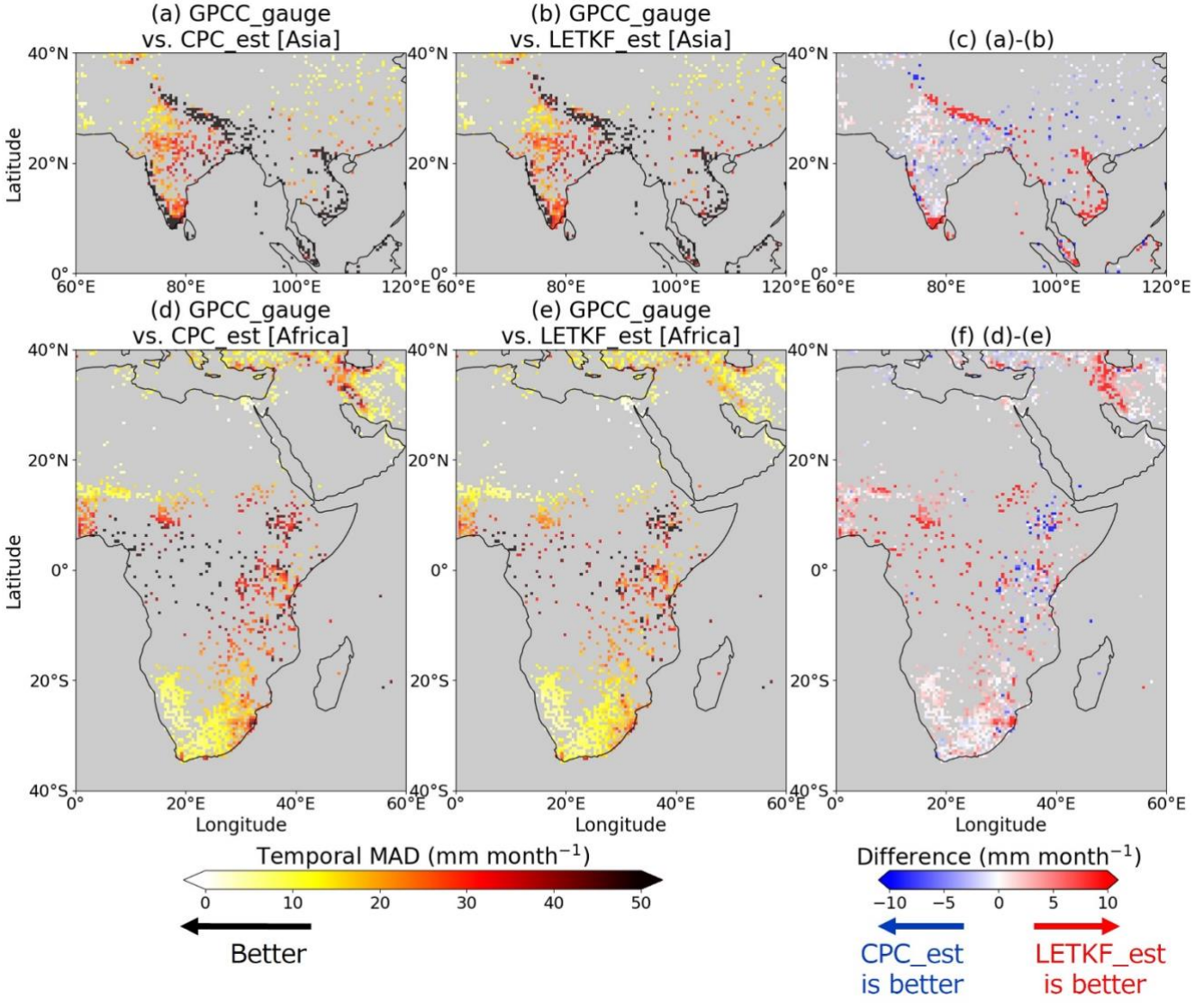

Figure 9: The temporal mean absolute difference (MAD) (mm month^{-1}) of (a, d) CPC_est and (b, e) LETKF_est computed against the monthly precipitation of GPCC_gauge at each pixel. Subplots (c, f) represent the differences (mm month^{-1}) between (a, d) and (b, e), respectively. Lighter colors in (a–b, d–e) indicate that the precipitation estimates are more similar to GPCC_gauge. Warm colors in (c, f) indicate that LETKF_est is more similar to GPCC_gauge than CPC_est, and cold colors indicate otherwise. The temporal MAD is computed only at pixels where more than 50 samples from GPCC_gauge are available, and the pixels are colored in gray if they do not match this condition.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

5 Conclusions

This study proposed a new estimation method for daily global precipitation fields from rain gauge observations using the algorithm of the LETKF in which the first guess and its error covariance are developed based on the precipitation from the reanalyzed precipitation of ERA5. Our findings can be summarized as follows.

Our estimates showed better agreements against rain gauge observations compared to the existing product of the NOAA CPC. Because we utilized the same rain gauge observations for the inputs of our estimation as those used for the NOAA CPC product, it is indicated that the proposed estimation method outperformed that of the NOAA CPC (i.e., the OI). Our proposed method took the advantage of constructing a dynamically guaranteed first guess and background error variance using reanalysis data for interpolating precipitation fields. Additionally, the method of this study was shown to be particularly beneficial for mountainous or rain-gauge-sparse regions.

There are some remaining limitations for this study that should be dealt with in the future. Firstly, our study has not applied any transformation based on probability distributions for the daily precipitation prior to the estimation, even though the precipitation variable is known to be less Gaussian. Many previous studies have pointed out that the analysis may not match the solution of the Bayesian estimation when we apply data assimilation based on minimum variance estimation on a state variable that is non-Gaussian, making it difficult to obtain the optimal analysis (e.g., Posselt and Bishop, 2012). This problem may occur significantly for regions where the precipitation amount is small, considering the fact that the ensemble used in the estimation of this study may contain many samples near 0.0 mm day^{-1} for such regions. Although the proposed method outperformed the OI in general, there is a possibility that the accuracy of the precipitation estimates will be further improved by applying the transformation methods such as the Gaussian transformation to the daily precipitation data (Lien et al., 2013; Kotsuki et al., 2017) in the future experiments. Another limitation is the lack of sites where validation can be performed in specific regions. For example, the density of rain gauges used in CPC_gauge is especially high in North America, making it difficult to perform validations against rain gauge observations independent from the observation inputs of the estimation (Fig. 2 b) in this region. On the other hand, both the rain gauges in CPC_gauge and other independent rain gauges used in GPCC_gauge are lacking in the central part of Australia and the Arabian Peninsula (Fig. 2 b). Therefore, there is a possibility that the validations performed in this study may be biased by the results of the regions with a large number of rain gauges independent from CPC_gauge.

Despite the limitations noted above, the present study succeeded in improving the accuracy of precipitation fields estimated from rain gauge observations, which will lead to a more effective use of these observations.

Author contributions

Y. Muto conducted all the experiments of this study, and S. Kotsuki developed the methodology of the study.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

Code availability

The code that supports the findings of this study is available from the corresponding author upon reasonable request.

Data availability

All of the data used for this study are publicly available data. In addition, all of the data and codes used in this study are stored for 5 years at Chiba University.

Competing interests

The authors have no competing interests to declare.

Acknowledgements

This study was partially supported by the Japan Aerospace Exploration Agency (JAXA) Precipitation Measuring Mission (PMM), the JSPS Grants in Aid for Scientific Research (JP21J11113), JSPS Kakenhi Grants (JP21H04571, JP21H05002, JP22K18821), and IAAR Research Support Program of Chiba University. The CPC Global Unified Gauge-Based Analysis of Daily Precipitation data are provided by the NOAA PSL, Boulder, Colorado, USA, from their website at https://psl.noaa.gov. Hersbach et al. (2023) is provided by the Copernicus Climate Change Service.

References

Barnes, S. L.: A technique for maximizing details in nu- merical weather map analysis, J. Appl. Meteor., 3, 396–409, doi: 10.1175/1520-0450(1964)003<0396:ATFMDI>2.0.CO;2, 1964.

Becker A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U. and Ziese, M.: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present, Earth Sys. Sci. Data, 5, 1, 71–99, doi: 10.5194/essd-5-71-2013, 2013.

Chen, M., P. Xie and J. E. Janowiak: Global land precipitation: A 50-yr monthly analysis based on gauge observations, J. of Hydrometeorol., 3, 249–266, doi: 10.1175/1525-7541(2002)003<0249:GLPAYM>2.0.CO;2, 2002.

Cressman, G. P.: An operational objective analysis system, Mon. Wea. Rev., 87, 367–374, doi: 10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2, 1959.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, C5, 143−162, doi: 10.1029/94JC00572, 1994.

Gandin, L. S.: Objective analysis of meteorological fields. Israel Program for Scientific Translations, 242 pp, 1965.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47, 2023. (Accessed on 17-03-2024)

Hamrud, M., Bonavita, M., and Isaksen, L.: EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation, Mon. Wea. Rev., 143, 4847–4864. doi: 10.1175/MWR-D-14-00333.1, 2015.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, PhysicaD: Nonlinear Phenom., 230, 1−2, 112−126, doi: 10.1016/j.physd.2006.11.008, 2007.

Ji, X., Li, Y., Luo, X., He, D., Guo, R., Wang, J., Bai, Y., Yue, C. and Liu, C.: Evaluation of bias correction methods for APHRODITE data to improve hydrologic simulation in a large Himalayan basin, Atmospheric Research, 242, 104964, doi: 10.1016/j.atmosres.2020.104964, 2020.

Kalman, R. E.: A new approach to linear filtering and prediction problems, J. of Basic Eng., 82, 1, 35−45, doi: 10.1115/1.3662552, 1960.

Kendall, M.: Rank correlation methods, Charles Griffin & Company Limited, 272 pp, 1948.

Kotsuki, S., Tanaka, K.: Uncertainties of precipitation products and their impacts on runoff estimates through hydrological land surface simulation in Southeast Asia, Hydrol. Res. Let., 7, 4, 79–84, doi: 10.3178/hrl.7.79, 2013.

Kotsuki, S., Miyoshi, T., Terasaki, K., Lien, G.-Y. and Kalnay, E.: Assimilating the global satellite mapping of precipitation data with the Nonhydrostatic Icosahedral Atmospheric Model (NICAM), J. Geophys. Res. Atmos., 122, 631–650, doi:10.1002/2016JD025355, 2017.

Kotsuki, S. and Bishop, C. H.: Implementing hybrid background error covariance into the LETKF with attenuation-based localization: Experiments with a simplified AGCM, Mon. Wea. Rev., 150, 283–302, doi: 10.1175/MWR-D-21-0174.1, 2022.

Kretschmer, M., Hunt, B. R. and Ott, E.: Data assimilation using a climatologically augmented local ensemble transform Kalman filter, Tellus A, 67, 26617, doi: 10.3402/tellusa.v67.26617, 2015.

Kubota, T., Aonashi, K., Ushio, T. Shige, S., Takayabu,Y. N., Kachi, M., Arai, Y. Tashima, T., Masaki, T., Kawamoto, N., Mega, T., Yamamoto, M. K. ,Hamada, A., Yamaji, M., Liu, G. and Oki, R.: Global Satellite Mapping of Precipitation (GSMaP) Products in the GPM Era. Satellite Precipitation Measurement. Advances in Global Change Research, Springer, 67, 355–373, doi: 10.1007/978-3-030-24568-9_20, 2020.

Kumar, P., Gairola, R. M., Kubota, T. and Kishtawal, C. M.: Hybrid assimilation of satellite rainfall product with high density gauge network to improve daily estimation: A case of Karnataka, India, J. of the Meteorol. Soc. of Japan, 99, 3, 741−763, doi: 10.2151/jmsj.2021-037, 2021.

Lien, G.-Y., Kalnay, E. and Miyoshi, T.: Effective assimilation of global precipitation: simulation experiments, Tellus A, 65, 19915, doi: 10.3402/tellusa.v65i0.19915, 2013.

https://doi.org/10.5194/egusphere-2024-960
Preprint. Discussion started: 23 April 2024

Mega, T., Ushio, T., Takahiro, M., Kubota, T., Kachi, M. and Oki, R.: Gauge-Adjusted Global Satellite Mapping of Precipitation, IEEE Trans. Geosci. Remote Sensing, 57, 1928–1935, doi: 10.1109/TGRS.2018.2870199 2019.

Posselt, D. J. and Bishop, C. H.: Nonlinear parameter estimation: Comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm, Mon. Wea. Rev., 140, 6, 1957−1974, doi: 10.1175/MWR-D-11-00242.1, 2012.

Pu, Z. and Kalnay, E.: Numerical weather prediction basics: Models, numerical methods, and data assimilation, in: Handbook of Hydrometeorological Ensemble Forecasting, edited by: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H. and Schaake J., Springer, Berlin, Heidelberg, 1–31, doi: 10.1007/978-3-642-40457-3_11-1, 2018.

Schneider, U., Hänsel, S., Finger, P., Rustemeier , E. and Ziese, M.: GPCC Full data monthly product Version 2022 at 0.5: Monthly land-surface precipitation from rain-gauges built on GTS-based and historical data, doi: 10.5676/DWD_GPCC/FD_M_V2022_050, 2022. (Accessed on 17-03-2024)

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., and Potthast, R.: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA), Q. J. R. Meteorol. Soc., 142, 1453–1472. doi: 10.1002/qj.2748, 2016.

Shepard, D.: A two dimensional interpolation function for irregularly spaced data, Proceedings of the 1968 ACM National Conf., 517–524, doi: 10.1145/800186.810616, 1968.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S. and Hsu, K.-L.: A review on global precipitation data sets: Data sources, estimation, and intercomparisons, Rev. Geophys., 56, 79–107, doi: 10.1002/2017RG000574, 2018.

Terasaki, K., Sawada, M. and Miyoshi, T.: Local ensemble transform Kalman filter experiments with the Nonhydrostatic Icosahedral Atmospheric Model NICAM, SOLA, 11, 23−26, doi: 10.2151/sola.2015-006, 2015.

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

Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N. and Kitoh, A.: APHRODITE: Constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges, BAMS, 93, 9, 1401−1415, doi: 10.1175/BAMS-D-11-00122.1, 2012.