# Estimating global precipitation fields by interpolating rain gauge observations using the local ensemble transform Kalman filter and reanalysis precipitation

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 of 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), a computationally efficient ensemble data assimilation method, 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 for 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 consistent first guess and its error variance using reanalysis data for interpolating precipitation fields. Furthermore, validations against independent rain gauge observations showed that our estimates are largely improved in mountainous or rain-gauge-sparse regions compared to the CPC estimates, indicating strong benefits of the proposed method for such regions.

## 1 Introduction

Improving the accuracy of global precipitation fields is crucial for predicting water-related disasters such as floods and droughts, long-term water resource management, and validations of forecasted precipitation by numerical weather prediction (NWP) models. Ground rain gauge observations play an essential role in estimating global precipitation fields, because they are considered to be more accurate relative to other estimates by 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 on rain gauge observations. Such global precipitation data are important not only as input data to analyze the hydrological water cycle, but as 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, although rain-gauge-based global precipitation data are especially important for periods when no or few satellite observations were available, the methodology to improve global precipitation fields by utilizing precise ground rain gauge observations is valuable even with the advancements in satellite observations and numerical weather forecasting.

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 in each grid point and the surrounding observations. Because the OI determines the weights of the first guess and observation 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 (Xie et al. 2007), allowing this product to be the rain-gauge-based global precipitation estimates with the highest spatiotemporal resolution ($0.5° × 0.5°$ pixel daily data) to the present day (Sun et al. 2018). However, CPC_est was reported to smooth extreme values especially in rain-gauge-sparse regions (Shen and Xiong, 2016), and hence a better interpolation method would be beneficial.

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 daily climatological error covariance by regarding NWP-based precipitation fields 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 historical 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 in 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.

## 2 Methods

### 2.1 Interpolation method

#### 2.1.1 Input Data

This study uses the rain gauge data utilized in CPC_est for the observation inputs for the interpolation. CPC_est is published as $0.5° × 0.5°$ pixel data, and the rain gauge data used in it are collected by NOAA CPC from approximately 30,000 stations from multiple data sources such as 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, NCARS, 2022). Although CPC_est defines the daily precipitation by local time, we assume that the daily precipitation in CPC_est represents the 24-hour precipitation from 00:00 UTC, given that open information on the local time used for each pixel is limited and inaccurate. We only use the precipitation at pixels in which more than one rain gauge station is included (hereafter, CPC_gauge) for the observation inputs in our estimation, and also assume that the rain gauge(s) is (are) located at the center of each pixel. Since CPC_est is estimated by also using the rain gauge observations of CPC_gauge and thereafter interpolating

the precipitation field using the OI (Xie et al. 2007), we can compare the interpolation methods of the CPC product and our study by comparing the precipitation estimates themselves.

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 0.25° × 0.25° gridded hourly data based on the Integrated Forecasting 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 (i.e., 24-hour precipitation from 00:00 UTC) 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 schematic image of the interpolation method of this study is shown in Fig. 1.

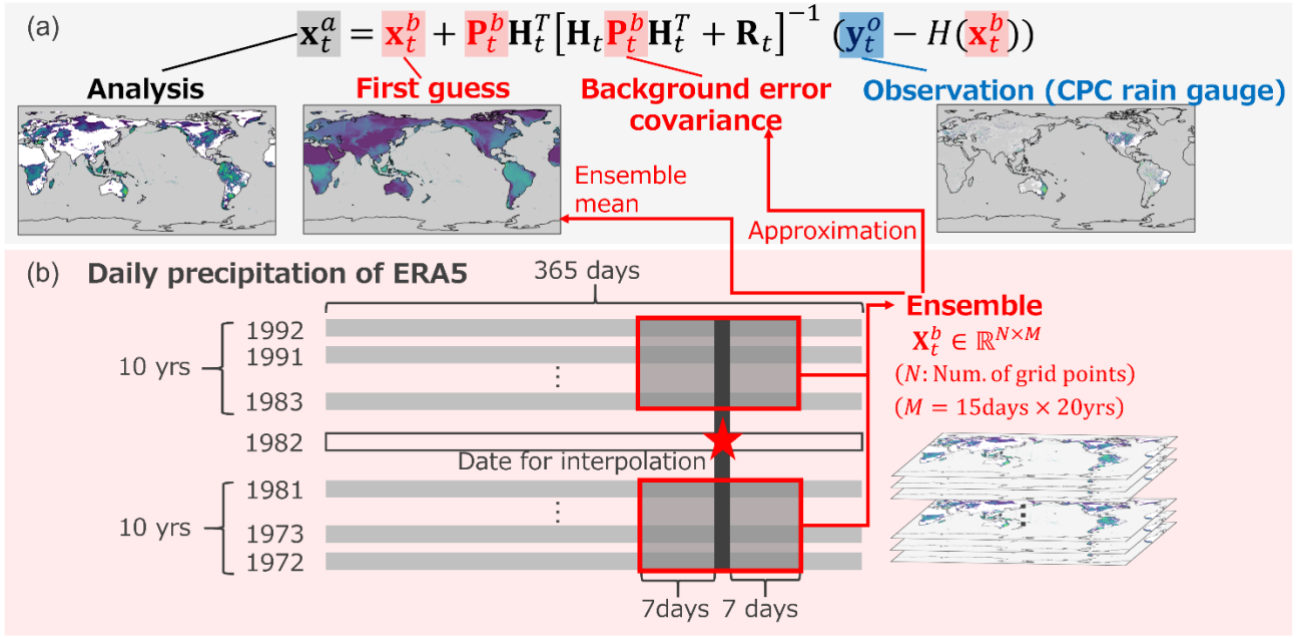



**Figure 1: The schematic image of (a) the interpolation method and (b) the construction of an ensemble in this study using ensemble data assimilation. The rain gauge observations from the CPC product are used for the observation $\mathbf{y}_t^o$. The ensemble $\mathbf{X}_t^b$ is obtained from the daily 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 $H_t(\ )$. 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 $\mathbf{x}_t^a$.**

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):

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

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. $\mathbf{P}_t^b \in \mathbb{R}^{N \times N}$ and $\mathbf{R}_t \in \mathbb{R}^{P \times P}$ represent the background and observation error covariance matrices. The scalars $N$ 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 $H_t(\ )$. Since we assume that the observation sites are located at the center of the 0.5-degree pixels, each observation site exactly corresponds to one 0.25-degree grid point of

110 the first guess. Hence, in our study, the observation operator $H_t()$ is simply a linear function which extracts the first guess data at grid points where the observation exists, and $\mathbf{H}_t$ is equivalent to $H_t()$.

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), based on Lien et al. (2016)'s suggestion about the effectiveness of logarithm transformation on precipitation variables:

$$\text{error variance} = \begin{cases} \log(2) & (y^o_{l,t} \leq 1.0 \text{ mm day}^{-1}) \\ \log(y^o_{l,t} + 1) & (y^o_{l,t} > 1.0 \text{ mm day}^{-1}) \end{cases}, \tag{2}$$

115

where log is the natural logarithm and $y^o_{l,t}$ denotes the observation at the $l$th pixel and $t$th time step from CPC_gauge. As described in Equation (2), there is a minimum limit ($\log(2)$) to the error variance, which prevents the inverse of $\mathbf{R}_t$ from diverging in Equations (7), (9) and (10) mentioned below. We also performed sensitivity experiments for a coefficient that multiplies the logarithm transformed value in Equation (2), and consequently the value 1.0 was selected as the coefficient (i.e.,

120 equivalent to placing no coefficient). An example of the spatial distribution of the error variances is shown in Appendix 1.

The first guess values of $\mathbf{x}^b_t$ and the background error covariance $\mathbf{P}^b_t$ 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, considering that CPC_est uses the 20-year average daily precipitation as the first guess for estimation (Xie et al., 2007). 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 $\mathbf{X}^b_t$ (cf.

125 Fig. 1 b) that represents the daily climatology of that date. 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 $\bar{\mathbf{x}}^b_t$ is given by the mean of the ensemble. Additionally, $\mathbf{P}^b_t$ is approximated by the ensemble (Evensen,1994), given by:

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

130

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

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

Ensemble data assimilation usually requires 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

135 and an observation site is $d$ km, the localization function $L(d)$ is expressed by the Gaussian function, which is widely used for localization in the LETKF algorithm (e.g., Miyoshi and Yamane, 2007):

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

where $\sigma$ 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 (i.e., the center of a 0.5° × 0.5° pixel), so that observations distant from that grid point

140 have less weights. The second row of Equation (5) truncates the observations where $d \geq 2\sqrt{10/3}\,\sigma$, based on Miyoshi et al. (2007). Here, we determine the value of $\sigma$ based on the method of Schraff et al. (2016), known as the Observation Number Limit technique. First, a certain distance $d^{ini}_{max}$ in km is set, followed by setting the maximum number of observation sites ($P^{max}_{loc}$) to be used for the estimation. Next, the localization scale $\sigma$ is determined by Equation (6):

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

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 explained in Appendix 2. Additionally, examples of $L(d)$ values with difference $\sigma$ values are shown in Appendix 3.

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):

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

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

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

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

where $\mathbf{I}$ denote the identity matrix, and 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 $\left( \widetilde{\mathbf{P}}_t^a \right)^{-1}$.

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 Validation methods

### 2.2.1 Data used for the validations

Rain gauge observation data from two precipitation products are used for 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. The use of such rain gauge data enables validation against rain gauge observations independent from those used in CPC_est. Here, we use the $0.5° \times 0.5°$ pixel data of the latest version of APHRODITE (V1101) covering Monsoon Asia (MA) (Fig. 2 a), where particularly dense rain gauge data other than those from the GTS are available in the APHRODITE product.

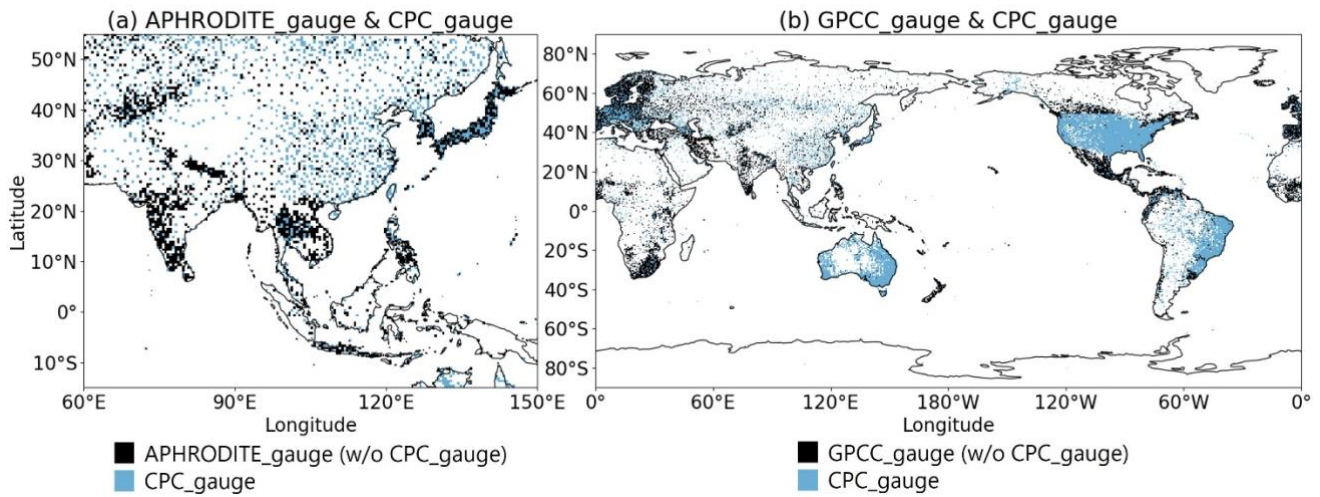



**Figure 2: An example of (a) the distribution of the daily rain gauge observations used in APHRODITE v1101 and the CPC product in Monsoon Asia (on Nov. 15th, 1988), and (b) the monthly rain gauge observations used in the GPCC FD product v2022 and the CPC product (in Nov., 1988). The black pixels include 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 data by the World Meteorological Organization or the collection of the Global Historical Climatology Network (Becker et al., 2013). Thus, although on a monthly basis, the GPCC provides rain gauge observations independent from CPC_gauge in a global scale (Fig. 2 b). In this study, we use the latest version of the $0.5° \times 0.5°$ pixel FD product (v2022) (Schneider et al., 2022).

The objective of this study is to improve the accuracy of rain-gauge-based precipitation fields in a global scale. Considering that the MA APHRODITE product in the MA has a limitation in area, and that the GPCC product has that in the temporal resolution, we perform validations against both of the data for more comprehensive evaluations.

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 $0.25° \times 0.25°$ gridded LETKF_est and ERA5 data are converted into $0.5° \times 0.5°$ pixel data so as to be equivalent to the spatial resolution of CPC_est, APHRODITE_gauge and GPCC_gauge. The details of the conversion method are described in Appendix 4.

### 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 that the APHRODITE precipitation underestimates annual, monthly and daily precipitation in South-east Asia (Kotsuki and Tanaka, 2013) and South Asia (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 $\tau_b$ (Kendall, 1948) is computed against the daily precipitation of APHRODITE_gauge for LETKF_est, CPC_est, and ERA5, respectively, using the data during the whole estimation period of this study (1981−1990). When $N_{aphro}$

205 is the number of APHRODITE_gauge pixels, and $(u_i, v_i)$ $(i = 1, ..., N_{aphro})$ are the pairs of daily precipitation data to be compared (i.e., the precipitation estimates and APHRODITE_gauge), $\tau_b$ is obtained by Equations (11) and (12):

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

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

where $A$ ($B$) represent the total number of cases in which the magnitude relationship of $u_j$ $(j = 1, ..., N_{aphro})$ and $u_k$ $(k = $
210 $j + 1, ..., N_{aphro})$ 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 $\tau_b$ closer to 1.0 (–1.0) indicates stronger positive (negative) correlation between the two types of data. Because the computation of $\tau_b$ neglects the samples with the completely same values in APHRODITE_gauge (or in the precipitation estimates), and because there are more than one no-rain cases in APHRODITE_gauge or the precipitation
215 estimates, it should be mentioned that $\tau_b$ cannot measure the similarity of no-rain cases between the two 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 below this value generally cannot be measured precisely by rain gauges.

**2.2.3 Validations against GPCC_gauge**

220 The spatial root mean square difference (RMSD), mean absolute difference (MAD) and Pearson's correlation coefficient (R) are computed for each month during the whole estimation period (1981−1990) against the monthly precipitation of GPCC_gauge for LETKF_est and CPC_est following Equations (13) to (15):

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

225
$$Spatial\ MAD_t = \frac{\sum_{i=1}^{N_{gpcc}} w_i \left|x_{gpcc\ i,t} - x_{est\ i,t}\right|}{\sum_{i=1}^{N_{gpcc}} w_i}, \tag{14}$$

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

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 $i$th pixel and $t$th 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 $t$th time step,
230 respectively. Here, $w_i = \cos(\theta_i)$ is the latitude-dependent weight of the $i$th pixel, where $\theta$ 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
235 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 1988 are 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

240   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, South America, and the Indochinese Peninsula. Precipitation areas can be seen around the Himalayas and the Zagros mountains

245   in LETKF_est, while not in CPC_est. In addition, the precipitation is generally weaker for LETKF_est than CPC_est.

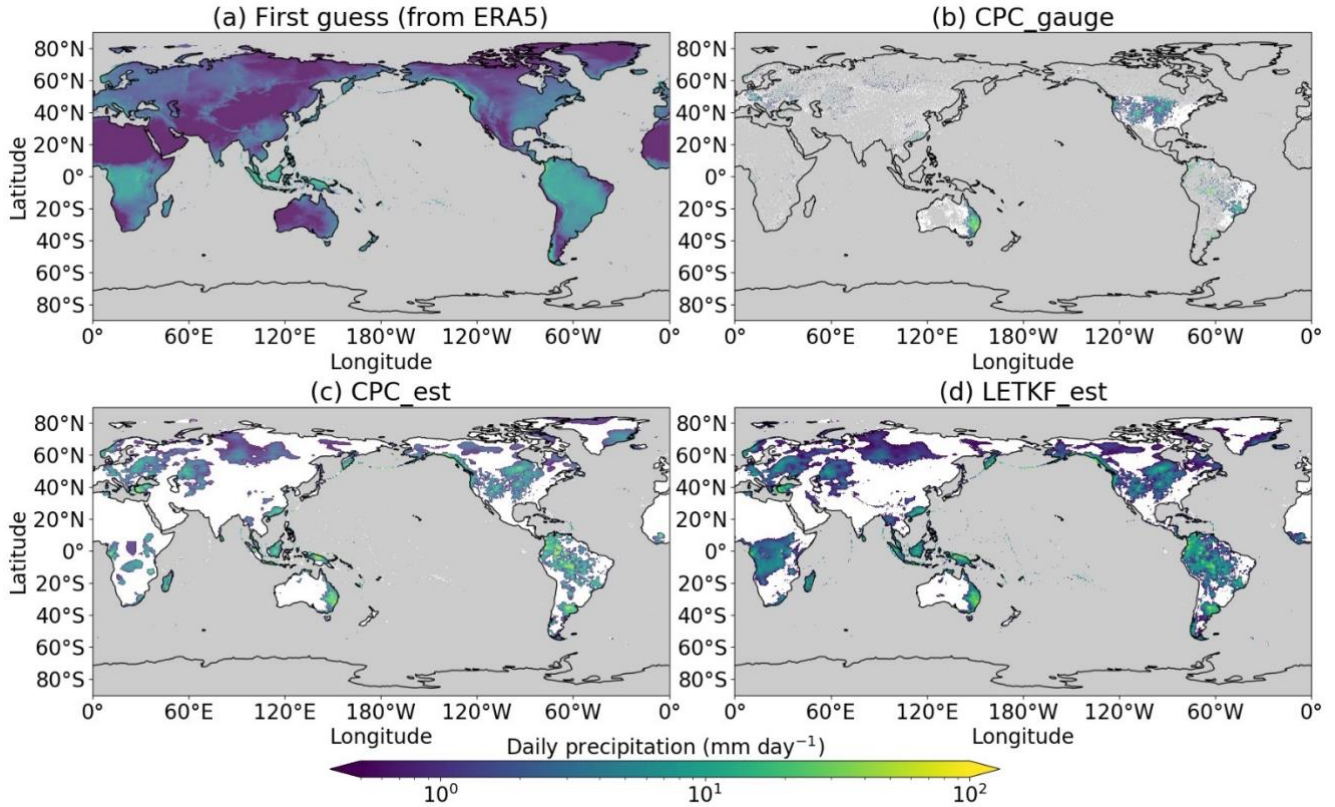



**Figure 3: An example of the precipitation fields (mm day$^{-1}$) for (a) the first guess used in our study, (b) the rain gauge observations of CPC_gauge, and the global precipitation estimates of (c) CPC_est and (d) LETKF_est (on Nov. 15th, 1988). Pixels on the ocean are colored in gray for all subplots,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

250   **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 $\tau_b$ value (described in Section 2.2.2) of LETKF_est computed against

255   APHRODITE_gauge is the highest (Fig. 4) with statistically significant differences at the P-value of 0.01, notwithstanding that LETKF_est was converted to 0.5˚ × 0.5˚ 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.

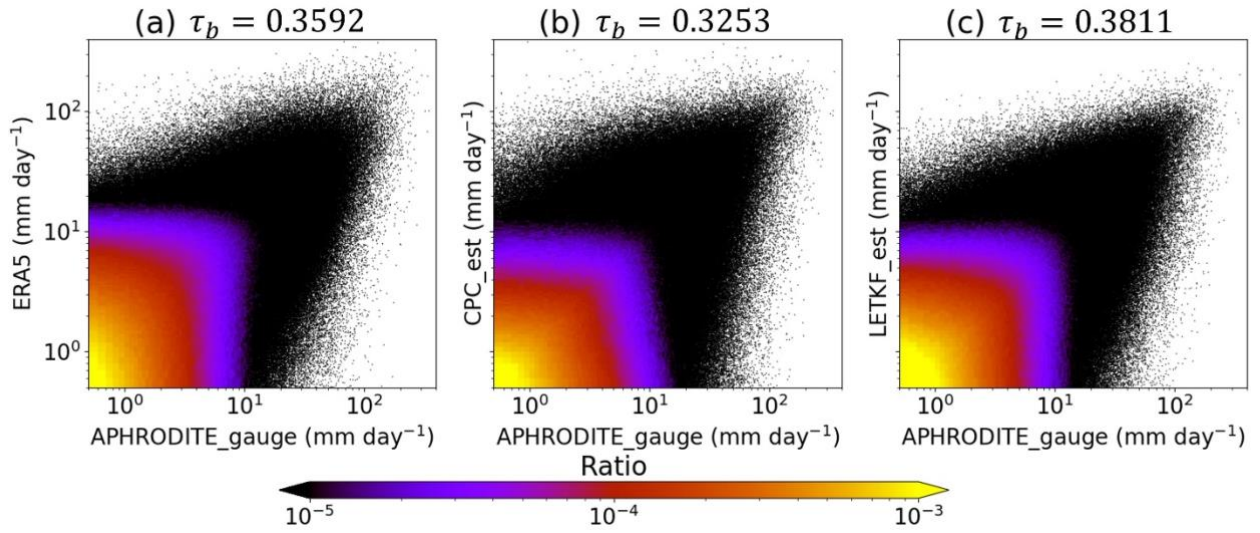



**Figure 4: Scatter plots comparing the daily precipitation (mm day$^{-1}$) of APHRODITE_gauge with that of (a) ERA5, (b) CPC_est and (c) LETKF_est. The colors represent the ratio of samples within each 0.1 mm day$^{-1}$ × 0.1 mm day$^{-1}$ bin in each 2-dimensional histogram. Kendall's rank correlation coefficient ($\tau_b$) 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 (described in Section 2.2.3) 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.

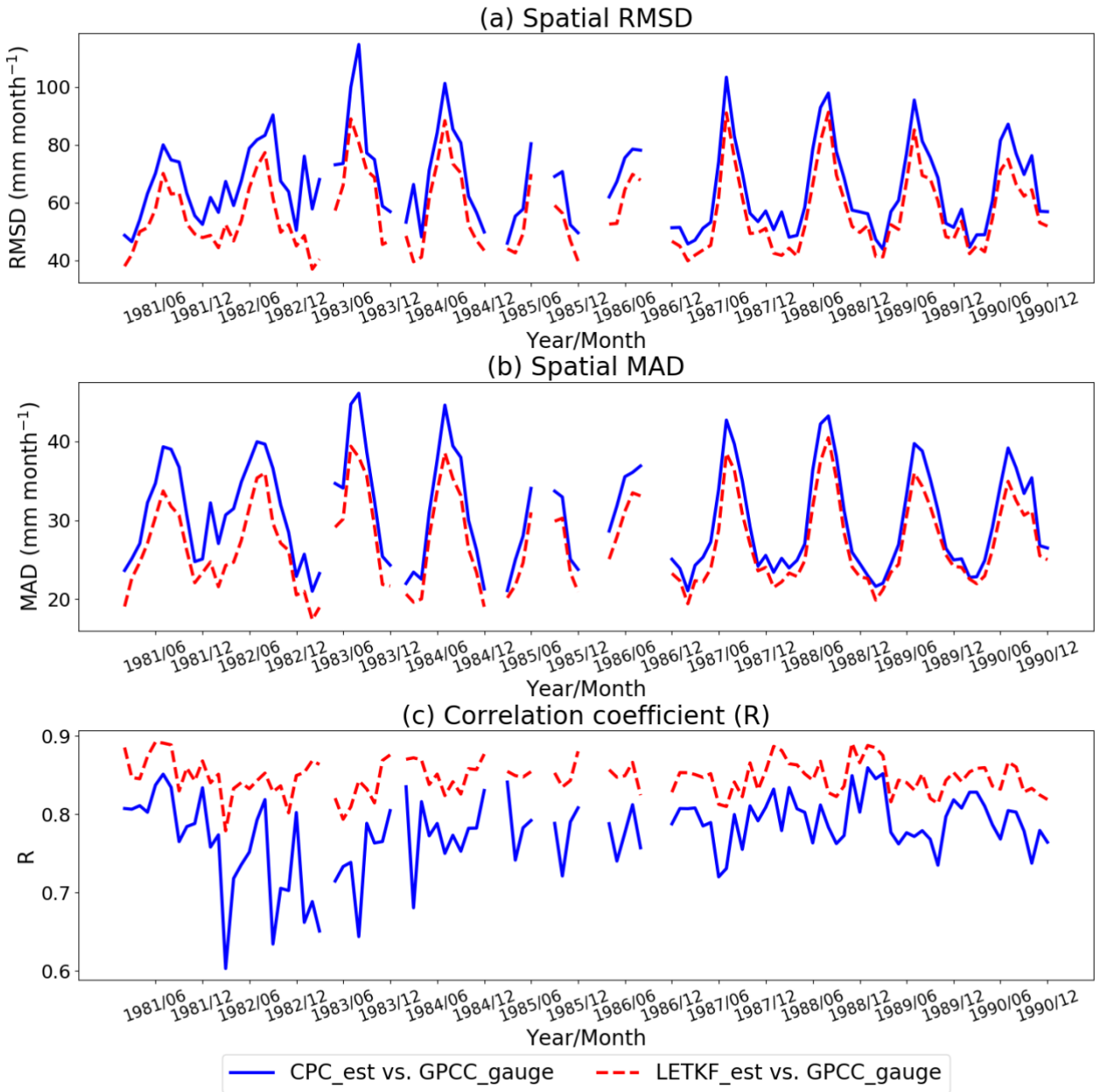



**Figure 5: The time series of (a) the spatial root mean square difference (RMSD; mm month⁻¹), (b) the spatial mean absolute difference (MAD; mm month⁻¹) and (c) Pearson's correlation coefficient (R), verified against the GPCC_gauge. The blue solid and red dashed lines represent the CPC_est and LETKF_est, 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).**

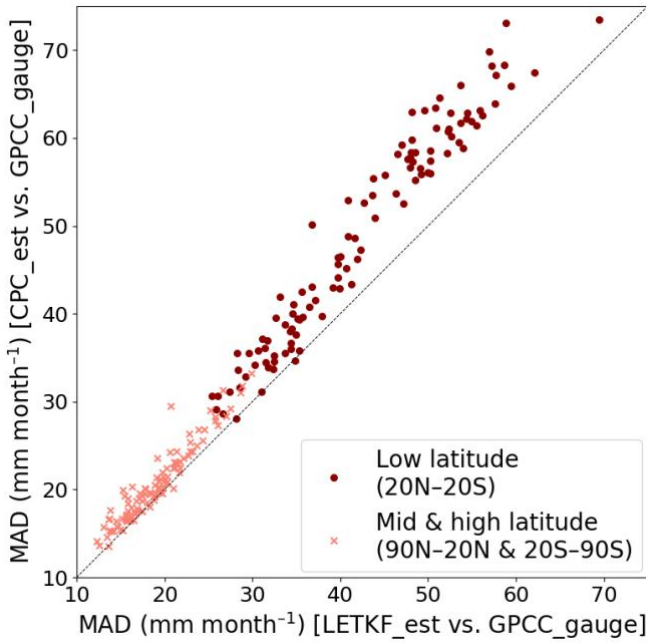



**Figure 6: Scatter plots comparing the spatial mean absolute difference (MAD; mm month$^{-1}$) of CPC_est and LETKF_est verified against the monthly precipitation of GPCC_gauge. Dark-red circles and light-red cross marks represent the low-latitude region (20˚N–20˚S) and mid- and high-latitude regions (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 consistent 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 first investigate the difference in South Asian precipitation, showing an example on an arbitrarily selected date in the Monsoon season.

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. 27th, 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) despite the lack of observation inputs in the surrounding area (Fig. 7 c), while CPC_est fails to do so (Fig. 7 e). Although the first guess of CPC_est is also adjusted considering orographic effects prior to interpolation by the OI (Xie et al., 2007), Fig. 7 e indicates that this adjustment would be insufficient. 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 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. It should be noted that orographic effects at a finer scale may be suboptimal in our estimation method, in which the rain gauge sites are assumed to locate at the center of 0.5˚ × 0.5˚ pixels. Furthermore, the performance of our proposed method may also differ if different reanalysis data is used, because reanalysis data with better quality would provide a better first guess and error covariance.

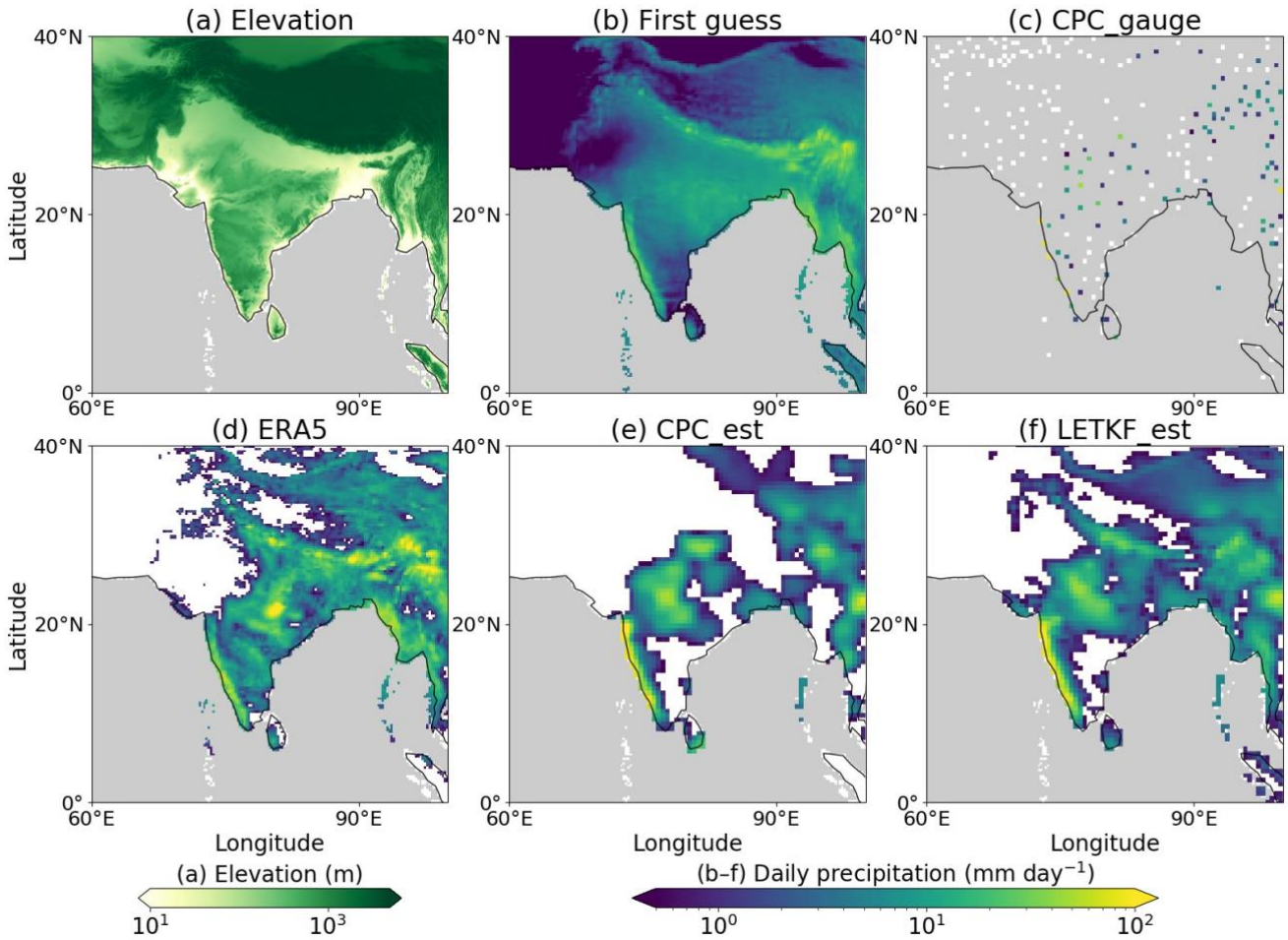



**Figure 7: (a) The elevation (m) and an example of (b) the first guess constructed in our study (mm day⁻¹), (c) the rain gauge observations of CPC_gauge (mm day⁻¹), and the global precipitation estimates (mm day⁻¹) of (d) ERA5, (e) CPC_est and (f) LETKF_est (on Jun. 27th, 1985) around India. Pixels 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.**

To investigate whether the precipitation of LETKF_est is more accurate than that of CPC_est around mountainous areas such as the Himalayas not only on a specific date but for the whole estimation period, Kendall's rank correlation coefficient ($\tau_b$) 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 are available during the whole estimation period (1981–1990) in MA. The results in Fig. 8 show that the $\tau_b$ 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 largely around this area during the estimation period.

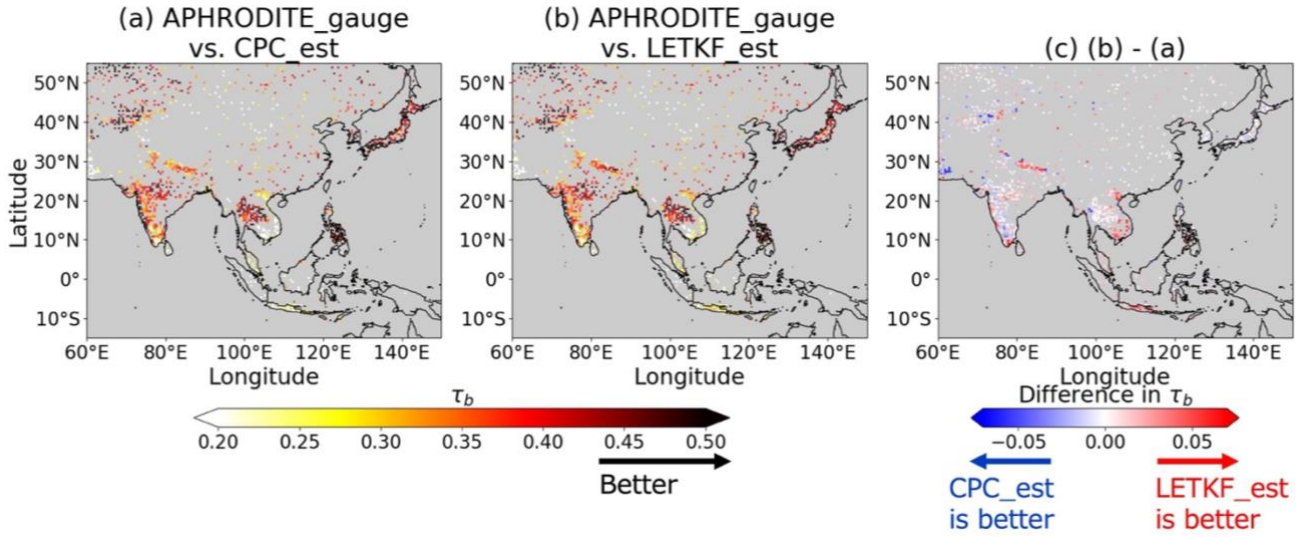



**Figure 8: Kendall's rank correlation coefficient ($\tau_b$) computed against the daily precipitation of APHRODITE_gauge for (a) CPC_est and (b) LETKF_est at each pixel. Subplot (c) represents the difference between (b) and (a). Darker colors in (a–b) indicate that the precipitation estimates are more 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. $\tau_b$ 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.**

Additionally, the temporal MAD values of LETKF_est and CPC_est are computed for the global area against the monthly precipitation of GPCC_gauge at 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 monthly time steps during the whole estimation period (1981–1990). The spatial MAD (described in Section 2.2.3) shows the similarity between two data in the global scale for each month, whereas the temporal MAD shows the similarity between two data for the whole period at each pixel.

Figure 9 depicts the temporal MAD computed in the Asian and African regions. The results for the global area are also shown in Appendix 5. 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 throughout the estimation period. 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 some regions in Southeast Asia (Fig. 9 c) or in 0˚S–20˚S 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. Moreover, the 10-year experiment of our study is completed in < 12 hours (i.e., < 12 seconds to estimate a daily global precipitation field) using 20 cores in the computer processing unit AMD EPYC Rome 7402, indicating the expected computational efficiency of the LETKF algorithm, as mentioned in Section 1. Since reanalysis data covers various variables other than precipitation, there is also a possibility that our proposed method is applicable to other variables such as soil moisture, depending on the accuracy, frequency and spatial density of its observations.

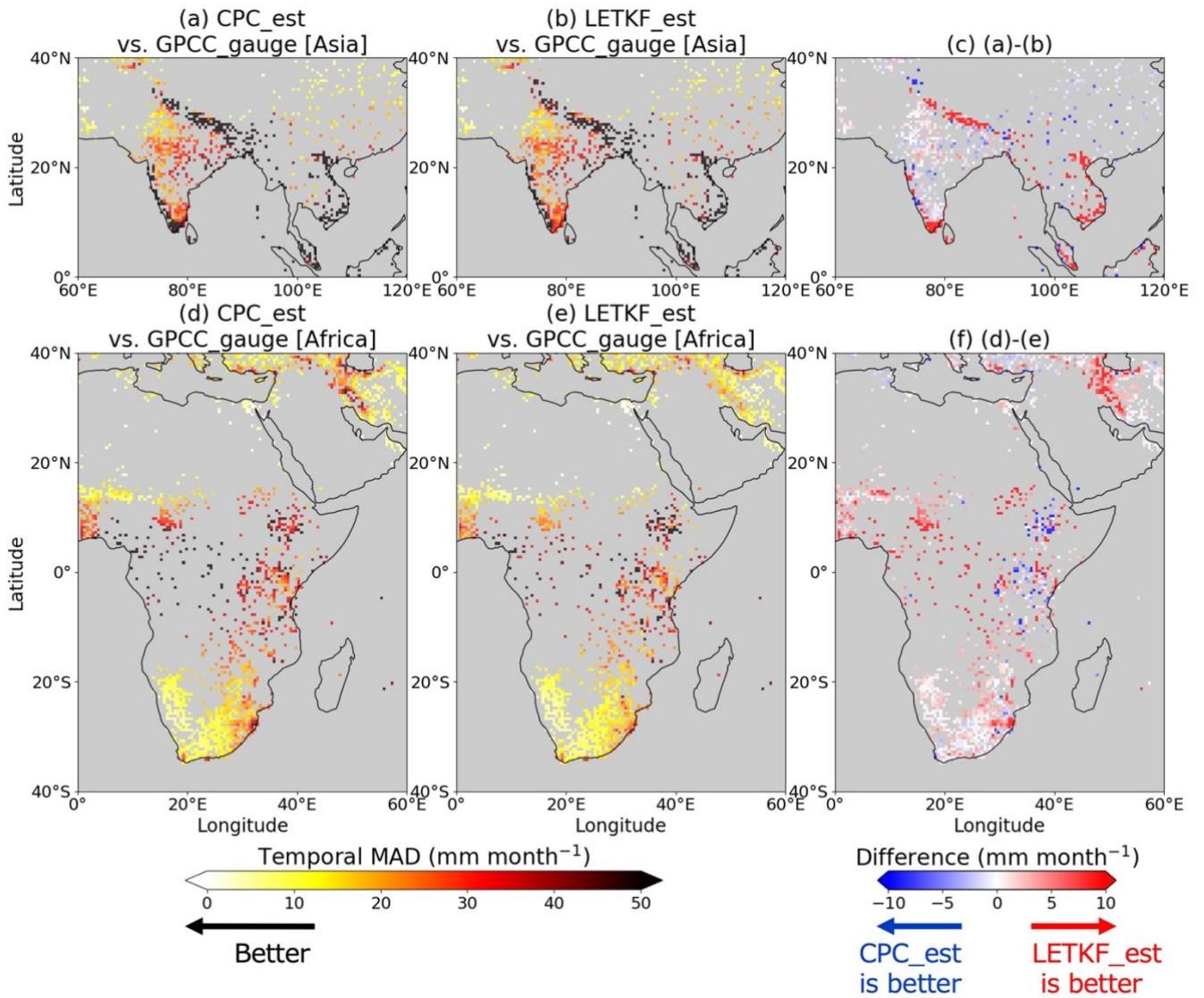



**Figure 9: The temporal mean absolute difference (MAD) (mm month⁻¹) of (a, d) CPC_est and (b, e) LETKF_est computed against the monthly precipitation of GPCC_gauge at each pixel. Subplots (c, f) represent the differences (mm month⁻¹) between (a, d) and (b, e), respectively. Lighter colors in (a–b, d–e) 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.**

There are some remaining limitations for this study that should be dealt with in the future. Firstly, our study has applied no transformation for the probability distributions of daily precipitation prior to the estimation, even though the precipitation variable can be less Gaussian. Many previous studies have pointed out that the analysis may not match the solution of the Bayesian estimation when data assimilation based on minimum variance estimation is applied to variables that are known to diverge from Gaussian, making it difficult to obtain the optimal analysis (e.g., Posselt and Bishop, 2012; Kotsuki et al., 2017). This problem may occur significantly for regions where the precipitation amount is small, considering the fact that the ensemble used in the estimation may contain many samples near 0.0 mm day⁻¹ for such regions. As such, although the proposed method outperformed the OI in general, there is a possibility that the accuracy of the precipitation estimates would be further improved by applying treatments on non-Gaussianity such as by the Gaussian transformation (Lien et al., 2013; Kotsuki et al., 2017) or the Gamma-Inverse-Gamma-Gaussian ensemble Kalman filtering (Bishop, 2016).

370   Another limitation is the lack of validation sites 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 sparse in the central part of Australia and the Arabian Peninsula (Fig. 2 b). Therefore, validations may be biased by the results of the regions with a large number of rain gauges

375   independent from CPC_gauge.

## 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. We succeeded in estimating the daily global precipitation fields with high computational

380   efficiency (i.e., < 12 seconds per day). 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, our results indicate that the proposed estimation method outperformed that of the NOAA CPC (i.e., the OI). Our proposed method took advantage of constructing a dynamically consistent first guess and background error variance

385   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 such as treatments on the less-Gaussian distribution of the precipitation variable, and the discrepancies among regions in the density of validation sites. Despite such limitations, the present study succeeded in improving the accuracy of precipitation fields estimated from rain gauge observations, which will

390   lead to a more effective use of spatially sparse rain-gauge observations.

**Author contributions**

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

**Code availability**

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

395   **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.

400   **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), JSPS Core-to-Core Program (JPJSCCA20220008) and IAAR Research Support Program and VL 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 numerical 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.

Bishop, C. H.: The GIGG-EnKF: ensemble Kalman filtering for highly skewed non-negative uncertainty distributions, Q. J. R. Meteorol. Soc., 142, 1395–1412. doi: 10.1002/qj.2742, 2016.

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.

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.

Lien, G.-Y., Miyoshi, T. and Kalnay E.: Assimilation of TRMM Multisatellite Precipitation Analysis with a low-resolution NCEP Global Forecasting System, Mon. Weather Rev., 144, 643–661, doi:10.1175/MWR-D-15-0149.1, 2016.

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.

Miyoshi, T. and Yamane, S.: Local ensemble transform Kalman filtering with an AGCM at a T159/L48 resolution, American Meteorol. Soc., 135, 3841–3861, doi: 10.1175/2007MWR1873.1, 2007.

Miyoshi, T., Yamane, S. and Enomoto, T.: Localizing the error covariance by physical distances within a local ensemble transform Kalman filter (LETKF), SOLA, 3, 89–92, doi:10.2151/sola.2007–023 2007.

National Center for Atmospheric Research Staff (Eds): The Climate Data Guide: CPC Unified Gauge-Based Analysis of Global Daily Precipitation, https://climatedataguide.ucar.edu/climate-data/cpc-unified-gauge-based-analysis-global-daily-precipitation (Accessed on 2024-07-10)

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.

Shen, Y. and Xiong, A.: Validation and comparison of a new gauge-based precipitation analysis over mainland China, International Journal of Climatology, 36, 252–265, doi: 10.1002/joc.4341, 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.

**Appendix 1: Spatial distribution of observation error variances in our study**

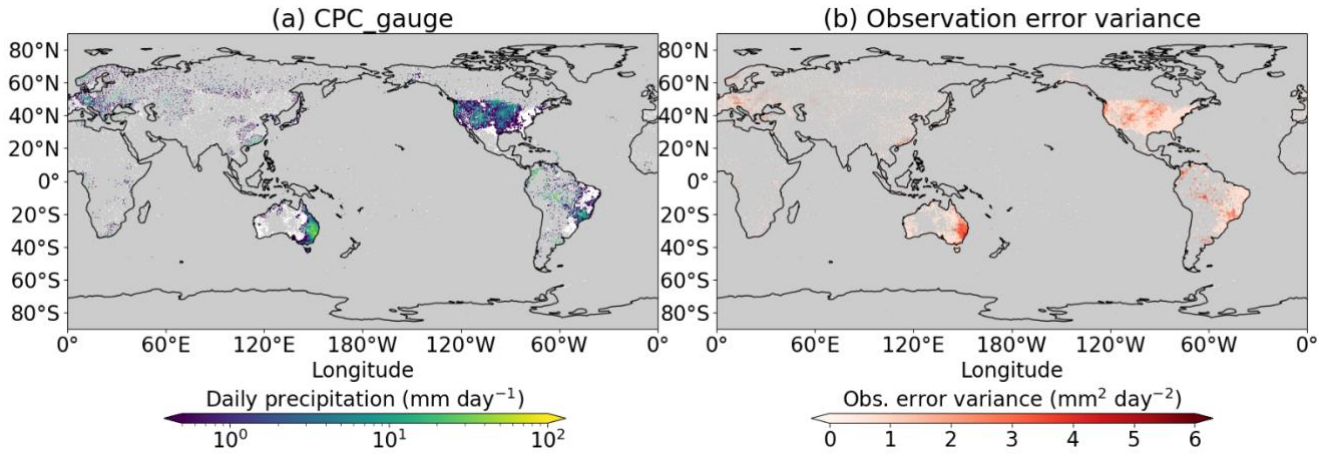



**Figure A1: An example of (a) the rain gauge observations of CPC_gauge (mm day$^{-1}$), and (b) the observation error variance (mm$^2$ day$^{-2}$) (on Nov. 15th, 1988). Pixels where no rain gauge observations are available are colored in gray for both subplots. Pixels are colored in white when the precipitation is < 0.5 mm day$^{-1}$ in Subplot (a).**

**Appendix 2: Sensitivity analysis of the localization parameters**

First, ten-year experiments from 1981 to 1990 were performed to estimate daily global precipitation fields using the methodology described in Section 2.1.2 with different combinations of the localization parameters $d_{max}^{ini}(= 500, 1000, 1500$ km) and $P_{loc}^{max}(= 5, 10, 20)$. Next, the validation against APHRODITE_gauge, which is described in Section 2.2.2 is performed for the precipitation estimates of each experiment. The results of the validations show that Kendall's rank correlation coefficient $\tau_b$ is the highest when $d_{max}^{ini} = 500$ km and $P_{loc}^{max} = 10$, followed by when $d_{max}^{ini} = 500$ km and $P_{loc}^{max} = 20$, and when $d_{max}^{ini} = 1000$ km and $P_{loc}^{max} = 10$ (Fig. A3). Because some grid points in Africa were found to have no observation point within a 500 km radius circle, values $d_{max}^{ini} = 1000$ km and $P_{loc}^{max} = 10$ were eventually selected for the localization parameters in the experiment described in the main text.

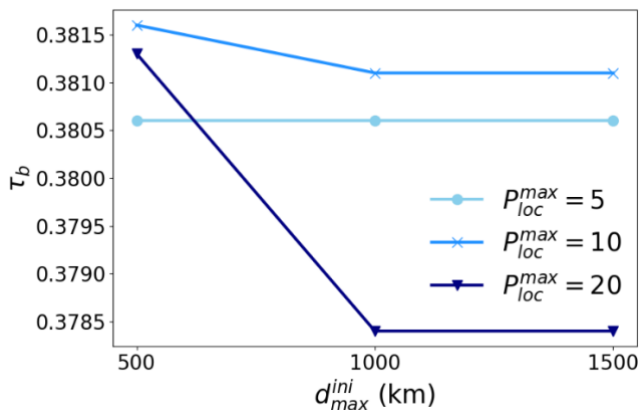
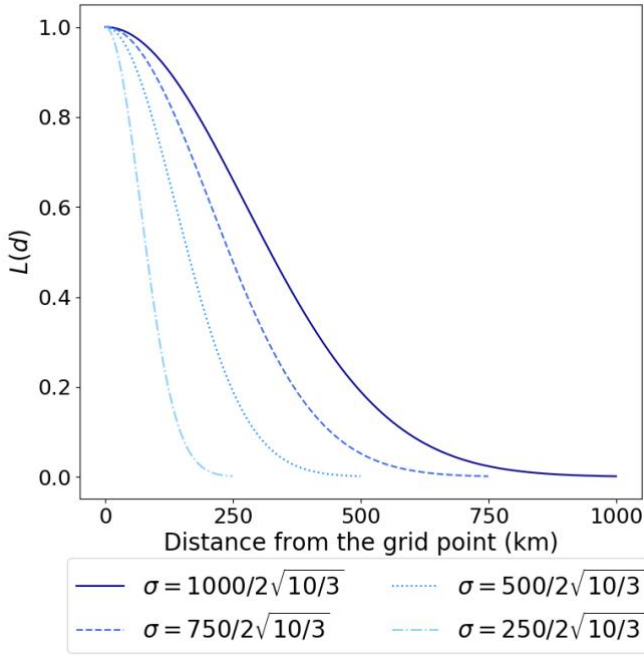



**Figure A2: Kendall's rank correlation coefficient computed against APHRODITE_gauge for different combinations of the localization parameters $d_{max}^{ini}(= 500, 1000, 1500$ km) and $P_{loc}^{max}(= 5, 10, 20)$.**

**Appendix 3: Change in the localization function depending on the distance of a grid point and an observation site**

510 **Figure A3: Localization function $L(d)$ depending on the distance of a grid point and an observation site when the localization scale $\sigma = \dfrac{1000}{2\sqrt{10/3}}, \dfrac{750}{2\sqrt{10/3}}, \dfrac{500}{2\sqrt{10/3}}, \dfrac{250}{2\sqrt{10/3}}$ (km).**

**Appendix 4: Method for converting 0.25° × 0.25° gridded ERA5 and LETKF_est data into 0.5° × 0.5° pixel data**

In our study, we assume that the daily precipitation of a 0.25° × 0.25° grid point represents that of a 0.25° × 0.25° pixel whose center is located at the original grid point. Thus, to convert the 0.25° × 0.25° gridded data into 0.5° × 0.5° pixel data, we compute

515 the weighted average of the daily precipitation of the 0.25° × 0.25° grid points inside of each 0.5° × 0.5° pixel, depending on the area ratio of the 0.25° × 0.25° pixels (Fig. A4). This method allows to conserve the total precipitation in the global area before and after the conversion.

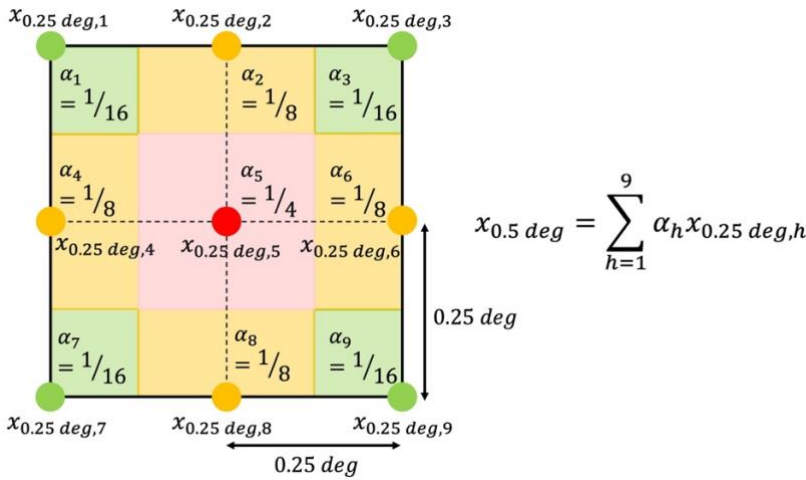



**Figure A4: Schematic image of the method for converting 0.25° × 0.25° gridded data (the precipitation data of the**

520 **colored plots) into 0.5° × 0.5° pixel data (the precipitation datum of the pixel surrounded by black lines). $x_{0.25\ deg,\ h}$ ($h = 1, ..., 9$) and $\alpha_h$ ($h = 1, ..., 9$) denote the daily precipitation and the weight of each 0.25° × 0.25° grid**

point. The daily precipitation of the $0.5° \times 0.5°$ pixel $x_{0.5\,deg}$ is computed by the weighted average of $x_{0.25\,deg,\,h}$ $(h = 1, ..., 9)$.

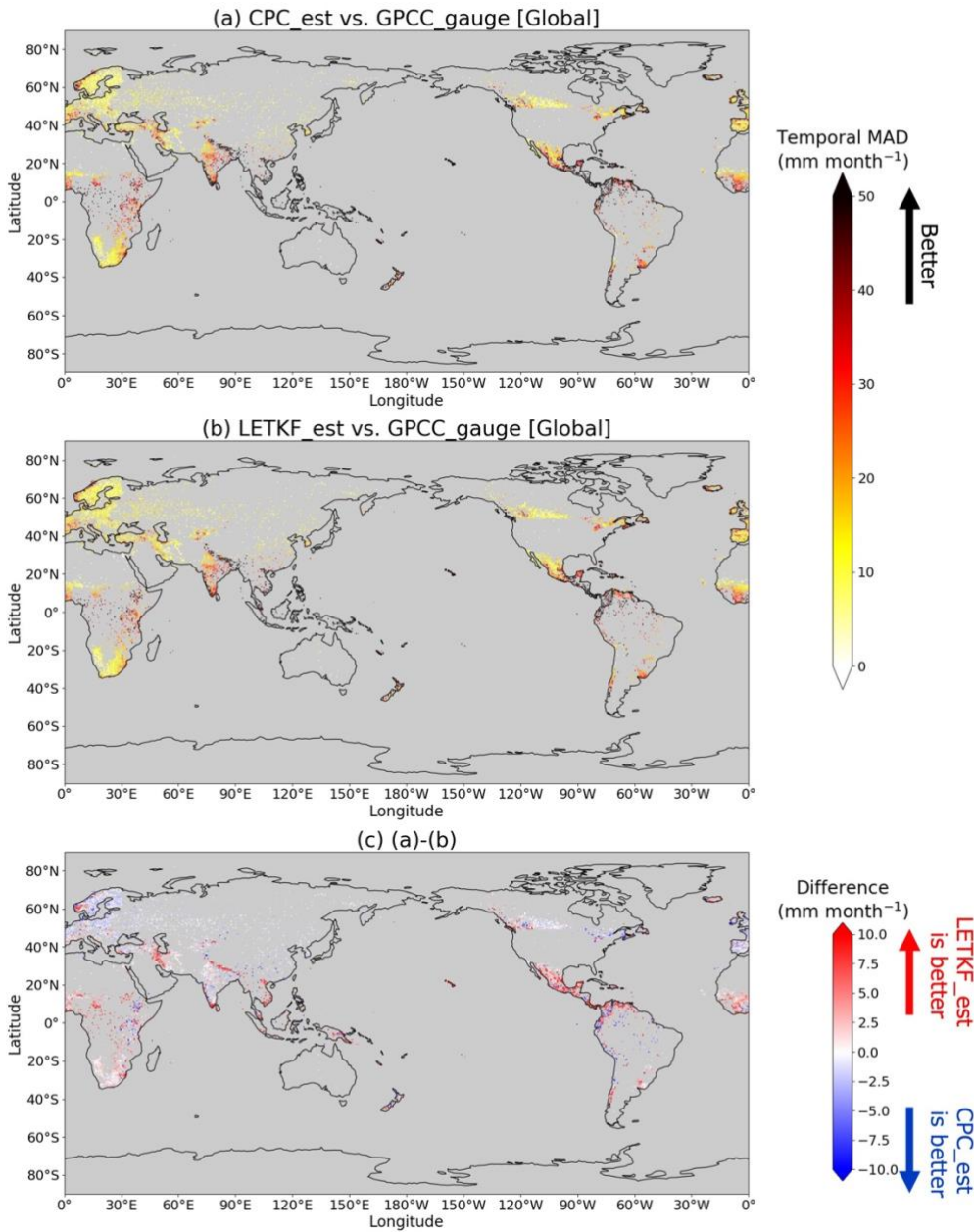
**Appendix 5: Spatial distribution of temporal MAD for the global area**

525

**Figure A5: The temporal mean absolute difference (MAD) (mm month⁻¹) of (a) CPC_est and (b) LETKF_est computed against the monthly precipitation of GPCC_gauge at each pixel in the global area. Subplot (c) represents the differences (mm month⁻¹) between (a) and (b). Lighter colors in (a–b) indicate that the precipitation estimates are more similar to GPCC_gauge. Warm colors in (c) indicate that LETKF_est is more similar to GPCC_gauge than CPC_est, and cold**
530 **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.**