




Intercomparison of precipitation datasets for summer precipitation characteristics over East Asia

Kwang-Yul Kim¹ · Jinwon Kim²  · Kyung-On Boo² · Sungbo Shim² · Youngmi Kim²

Received: 6 September 2017 / Accepted: 13 June 2018 / Published online: 23 June 2018
© The Author(s) 2018

Abstract

Precipitation data in the Global Precipitation Climatology Project (GPCP) and in four reanalysis datasets, ERA-Interim, MERRA, NCEP/NCAR, and JRA, are compared against the CPC Merged Precipitation (CMAP) in the cyclostationary empirical orthogonal function (CSEOF) space to evaluate these datasets in representing the summer precipitation characteristics over East Asia. CSEOF analysis is applied to each dataset, and regression analysis is performed in the CSEOF space with the CMAP data as the target. The regression analysis establishes one-to-one correspondence between the CSEOF loading vectors of the target variable and those of the predictors, i.e., GPCP and the four reanalysis datasets. The loading vectors of the GPCP data coincide almost exactly with those of the CMAP data, i.e., the two observation-based precipitation datasets represent practically identical summer precipitation characteristics over East Asia. The reanalysis datasets also reproduce the first five CSEOF modes reasonably; however, performance of NCEP/NCAR is notably lower than others. The re-constructed precipitation using the first five regressed CSEOF modes of the reanalysis datasets are well correlated with that of the CMAP data with reasonably large correlation coefficients, suggesting that these reanalysis precipitation products reliably simulate the major summer precipitation characteristics in East Asia. All of the four reanalysis products commonly show noticeable errors in representing the summer rainfall over the mid-latitude ocean to the south of Japan, the tropical western Pacific, tropical/subtropical regions including the Indochina Peninsula, India, the Maritime Continent, and regions of complex terrain especially those characterized by strong orographic slopes around the Tibetan Plateau. The errors over the regions of complex orography and coastal lines may be partially due to the inability of reanalysis models in simulating the effects of complex terrain and the lack of observations in these sparsely populated regions.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00382-018-4303-3>) contains supplementary material, which is available to authorized users.

✉ Jinwon Kim
jkim@atmos.ucla.edu
Kwang-Yul Kim
kwang56@snu.ac.kr
Kyung-On Boo
kyungon@korea.kr
Sungbo Shim
sbshim82@korea.kr
Youngmi Kim
kymi19@korea.kr

¹ School of Earth and Environmental Sciences, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

² National Institute of Meteorological Sciences, Korean Meteorological Administration, Seogwipo-si, Jeju-do, Republic of Korea

1 Introduction

Observation-based and reanalysis precipitation datasets play critical roles in climate research. For their importance, various research groups and operational centers around the world have recently introduced a number of precipitation analysis datasets based on observations from various platforms such as weather stations and satellite-borne sensors as well as reanalysis precipitation products based on model-assisted gridded assimilations. These datasets are great assets for advancing climate research via improved understanding of the dynamical and physical processes that shape the climate system (e.g., Kim et al. 2018; Neena et al. 2016) as well as by evaluating climate models for model improvement and bias corrections, just to name a few (Kim et al. 2013; Whitehall et al. 2012). Having multiple observational and reanalysis datasets available for climate research, accuracy of individual datasets or the spread among multiple datasets of the same field have become an important concern.

Previous studies (Prakash et al. 2014; Kim et al. 2015a; Kim and Park 2016) found that precipitation characteristics in observation-based gridded datasets often vary widely among them. This is an important concern because model evaluation against observational and/or reanalysis data is critical for model developments/improvements and in the application of climate model data to impact assessments via bias correction (Kim et al. 2015a), for example.

Because of the concern, earlier studies attempted to assess, quantify and understand the characteristics of uncertainties in observation-based gridded precipitation analysis datasets. Prakash et al. (2014), for example, found that observation-based precipitation analysis datasets yield varying summer monsoon rainfall climatology over India and that the inter-dataset differences vary notably according to geography. Kim et al. (2015a) revealed that the spread among observation-based gridded precipitation analyses varies seasonally as well as regionally. Kim and Park (2016) also showed that the spread among gridded precipitation datasets over East Asia varies according to statistical properties, i.e., smallest in seasonal means and largest in long-term trends, in addition to geography and season. Such differences arise from various sources involved in analysis processes such as input data quality control, observation density, analysis schemes, and, for reanalysis datasets, the differences in model formulations (Legates and Willmott 1990; Kim et al. 2015a; Kim and Park 2016). In most cases, it is difficult to trace and/or quantify the differences, not to mention to identify the most accurate dataset. In such cases, intercomparison of multiple datasets is one of useful ways of understanding the uncertainties in currently available observation-based datasets.

Most of previous data intercomparison studies dealt with the first two statistical moments, means and variances (e.g., Guirguis and Avissar 2008; Prakash et al. 2014; Kim et al. 2015a). In this approach, the regional and/or temporal variations of the differences between datasets can be assessed. On the other hand, it is difficult to assess differences in the spatial and temporal characteristics using only the first two moment statistics. The most critical limitation of this so-called “pedestrian” approach is that statistical differences between two datasets shed only limited insight into how these datasets differ from each other in terms of physics and dynamics. A dataset represents variations produced as a consequence of interactions of multiple physical and/or dynamical processes in both space and time. It would be highly beneficial if individual physical and/or dynamical processes in a dataset can be separated and evaluated separately. The cyclostationary empirical orthogonal function (CSEOF) technique is an ideal tool for decomposing a dataset into physically and dynamically distinct modes of spatiotemporal evolutions. Each CSEOF loading vector describes a temporal evolution of spatial patterns tied to a distinct physical and/

or dynamical mechanism. The loading vectors are mutually orthogonal, and the principal component (amplitude) time series are also uncorrelated (often nearly independent) in time. For details of the CSEOF analysis, readers are referred to Kim et al. (1996, 2015b) and Kim and North (1997). The essence of the approach used in this study is to evaluate individual CSEOF modes in order to assess how well reanalysis models perform in reproducing them. This approach allows us to evaluate a dataset in terms of the accuracy of physical and dynamical processes that produced the field in the dataset.

This study employs a recently introduced methodology based on the CSEOF analysis (Kim et al. 1996, 2015a; Kim and North 1997) to assess the accuracy of five precipitation products including the observation-based Global Precipitation Climatology Project (GPCP) and four reanalysis datasets against the observation-based CPC Merged Analysis of Precipitation (CMAP) dataset. Section 2 presents the methodology of analysis and the datasets used in this study. Results of analysis are presented in Sect. 3, followed by discussion and concluding remarks in Sect. 4.

2 Datasets and method of analysis

2.1 Datasets

This study analyzes six precipitation datasets for the 37-year period 1979–2015. Two of these datasets, CMAP (Xie and Arkin 1997) and GPCP (Huffman et al. 1997, 2009; Adler et al. 2003) are based on observations from surface stations and satellite retrievals. The spatial and temporal resolutions of these datasets are presented in Table 1. We arbitrarily select the CMAP data as the reference data against which the remaining five datasets are evaluated. Four reanalysis precipitation datasets including the ERA-Interim reanalysis (Dee et al. 2011), the NCEP/NCAR reanalysis (Kalnay et al. 1996), the Modern-Era Retrospective analysis for Research and Application (MERRA; Rienecker et al. 2011), and the Japanese 55-year Reanalysis (JRA; Kobayashi et al. 2015;

Table 1 The horizontal and temporal resolutions of the datasets employed in this study

	Horizontal resolution	Temporal resolution
CMAP	2.5° × 2.5°	5 days
GPCP	1.0° × 1.0°	5 days
ERA-Interim	0.75° × 0.75°	6 h
NCEP/NCAR	2.5° × 2.5°	6 h
MERRA	0.5° × 0.66°	6 h
JAR55	1.25° × 1.25°	6 h

Harada et al. 2016) are also evaluated against the CMAP data. Since the CMAP data are available at 5-day resolutions, the four reanalysis datasets are converted into pentad resolutions before the analysis. The native spatial resolutions of these six datasets also differ from each other. For consistency, all of these datasets are also interpolated to a common grid of $2.5^\circ \times 2.5^\circ$ resolutions for analyses. The re-scaled datasets preserve the fundamental precipitation characteristics of the original data (e.g., Guirguis and Avissar 2008). Other gauge-based daily observational datasets such as APHRODITE (Yatagai et al. 2012) that has been widely used in Asian monsoon and hydrology studies (e.g., Preethi et al. 2017) are also available, but not included in this study because they cover only land areas.

2.2 CSEOF analysis

In order to compare precipitation characteristics represented in each dataset, CSEOF analysis (Kim et al. 1996, 2015b; Kim and North 1997) is performed on all six datasets. This study is focused on the summer precipitation over the East Asian region of 90°E – 180°E and Equator– 60°N for the 120-day period (May 13–September 9). The 24-pentad summer period is selected in each year to describe evolutions of precipitation field during summer. CSEOF analysis decomposes space–time data as

$$T(r, t) = \sum_n B_n(r, t)T_n(t), \quad B_n(r, t) = B_n(r, t + d), \tag{1}$$

where r is location in space, t is time, n is the mode number, $d = 24$ pentads (i.e., 120 days) is the nested period, $B_n(r, t)$ are the cyclostationary loading vectors (CSLVs), and are the principal component (PC) time series for mode n . Each CSLV represents the temporal evolution of spatial patterns for the summer, and is orthogonal to others in space and time. Each PC time series describes long-term variations of the amplitude of the corresponding CSLV, and is uncorrelated with (often independent of) others. Conventional EOF analysis is a special case of CSEOF analysis with the nested period $d = 1$. Thus, each EOF loading vector consists of one spatial pattern and as a consequence is unable to depict an evolution of spatial patterns.

CSEOF analysis is conducted on another dataset $P(r, t)$ as in (2):

$$P(r, t) = \sum_n C_n(r, t)P_n(t), \quad C_n(r, t) = C_n(r, t + d). \tag{2}$$

One-to-one correspondence between CSEOFs of the two sets in (1) and (2) is not necessarily established. That is, $T_n(r, t) \neq P_n(r, t)$ in general, and $C_n(r, t)$ may represent different physical and/or dynamical mechanisms that cannot be compared directly against $B_n(r, t)$. In order to evaluate if and how well the two different datasets describe an identical

physical process, a regression analysis in CSEOF space should be performed.

2.3 Regression analysis in CSEOF space

Let us call $T(r, t)$ a target variable and $P(r, t)$ a predictor variable. Then, regression analysis in CSEOF space proceeds in the following manner:

$$\text{Step 1: } T_n(t) = \sum_{m=1}^M \alpha_m^{(n)} P_m(t) + \varepsilon^{(n)}(t), \tag{3}$$

$$\text{Step 2: } C_n^{(reg)}(r, t) = \sum_{m=1}^M \alpha_m^{(n)} C_m(r, t). \tag{4}$$

Here, (3) is multiple regression of a target PC time series in terms of the predictor PC time series. A new CSLV, called the regressed CSLV, can be found by a linear superposition of the predictor CSLVs using the regression coefficients $\{\alpha_m^{(n)}\}$. The parameter M is the number of predictor CSEOF modes used in the regression analysis in CSEOF space; $M = 20$ in this study. The predictor variable, then, can be rewritten as

$$P(r, t) = \sum_n C_n^{(reg)}(r, t)T_n(t). \tag{5}$$

Thus, the target and predictor variables together can be written as

$$\{T(r, t), P(r, t)\} = \sum_n \{B_n(r, t), C_n^{(reg)}(r, t)\}T_n(t). \tag{6}$$

The terms in the curly brackets represent an identical physical process manifested in the two different datasets. Thus, a comparison of $B_n(r, t)$ with $C_n^{(reg)}(r, t)$ reveals how similar (or dissimilar) is the rendition of a physical process in the two datasets. We can carry out regression analysis in CSEOF space for multiple predictor variables to write

$$Data(r, t) = \sum_n \{B_n(r, t), C_n^{(reg)}(r, t), \dots, Z_n^{(reg)}(r, t)\}T_n(t) \tag{7}$$

to compare multiple datasets against the target dataset.

3 Results

3.1 Raw statistics

The first two statistical moments, means and standard deviations, of the summer precipitation in the six datasets are compared in Fig. 1a, b. All the six datasets show similar spatial patterns and magnitudes of the first two moments.

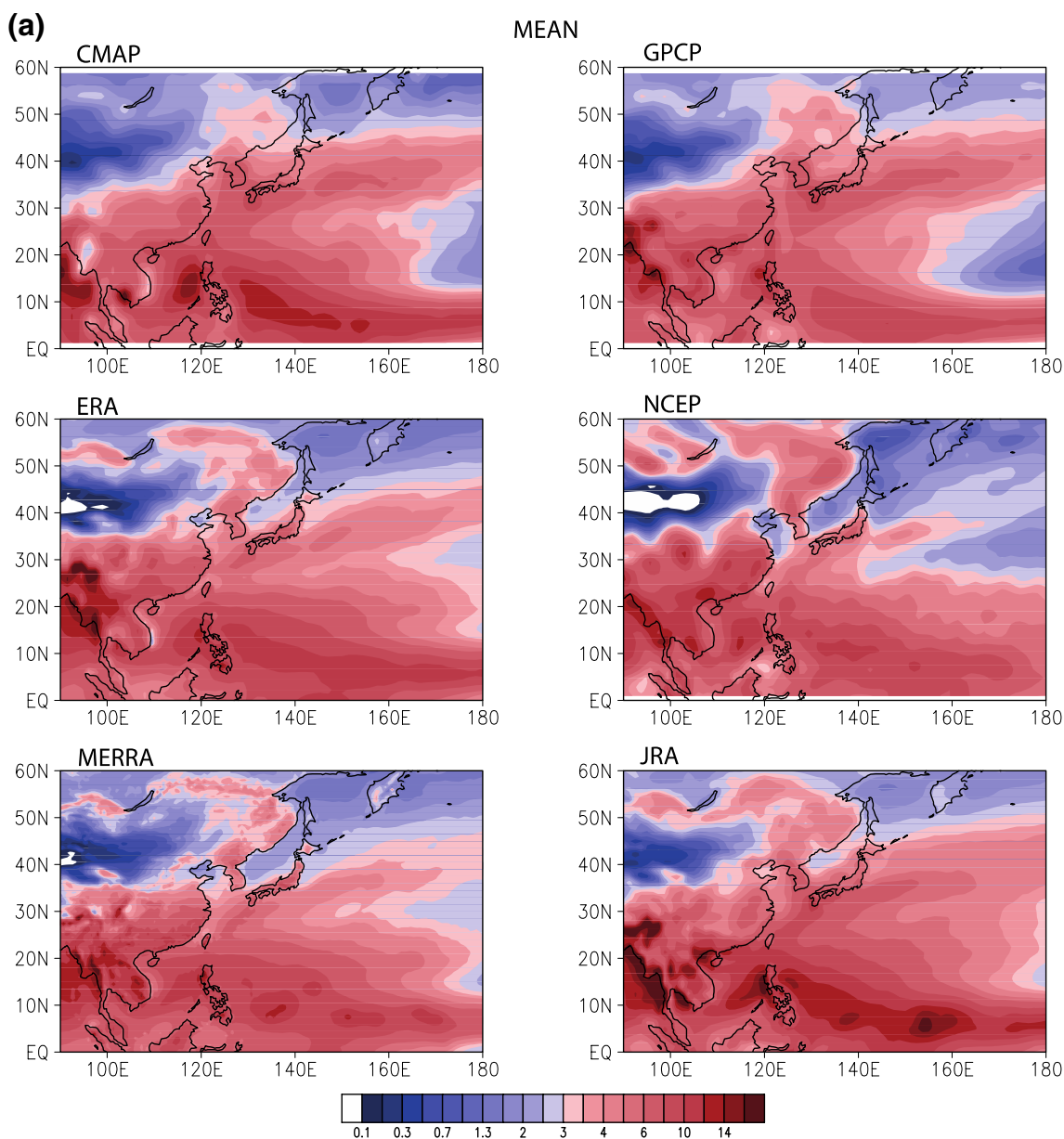


Fig. 1 a Summer-mean (May 13–September 9) precipitation climatology of the six datasets for 1979–2015. **b** Standard deviation of summer (May 13–September 9) precipitation of the six datasets for 1979–2015

Differences in the detailed structures between these datasets are also clear. Between the two observation-based datasets, CMAP shows slightly larger means and standard deviations than GPCP, most noticeably over the tropical western Pacific Ocean and off the west coast of the Indochina Peninsula (Table 2). Over the continental areas, on the other hand, GPCP data shows larger means and standard deviations (Table 2). The discrepancy between these two observation-based precipitation datasets could have originated from the differences in analysis procedures such as input datasets, pre-analysis data quality control, and analysis methodology (Xie and Arkin 1995; Vogel 2013; Kim et al. 2015a).

Previous studies (e.g., Kim et al. 2015a; Kim and Park 2016) showed that the differences between observation-based gridded precipitation datasets vary widely according to regions, seasons, precipitation characteristics and/or statistical properties. Overall, the long-term means and the temporal standard deviation over the 37 summers, a surrogate for the seasonal cycle and its interannual variability respectively, of the GPCP data agree closely with the CMAP data (Table 2).

The reanalysis precipitation datasets except JRA show systematically larger (smaller) means over the continental (ocean) region compared to the two observation-based products (Fig. 1a). Compared to CMAP, all reanalysis

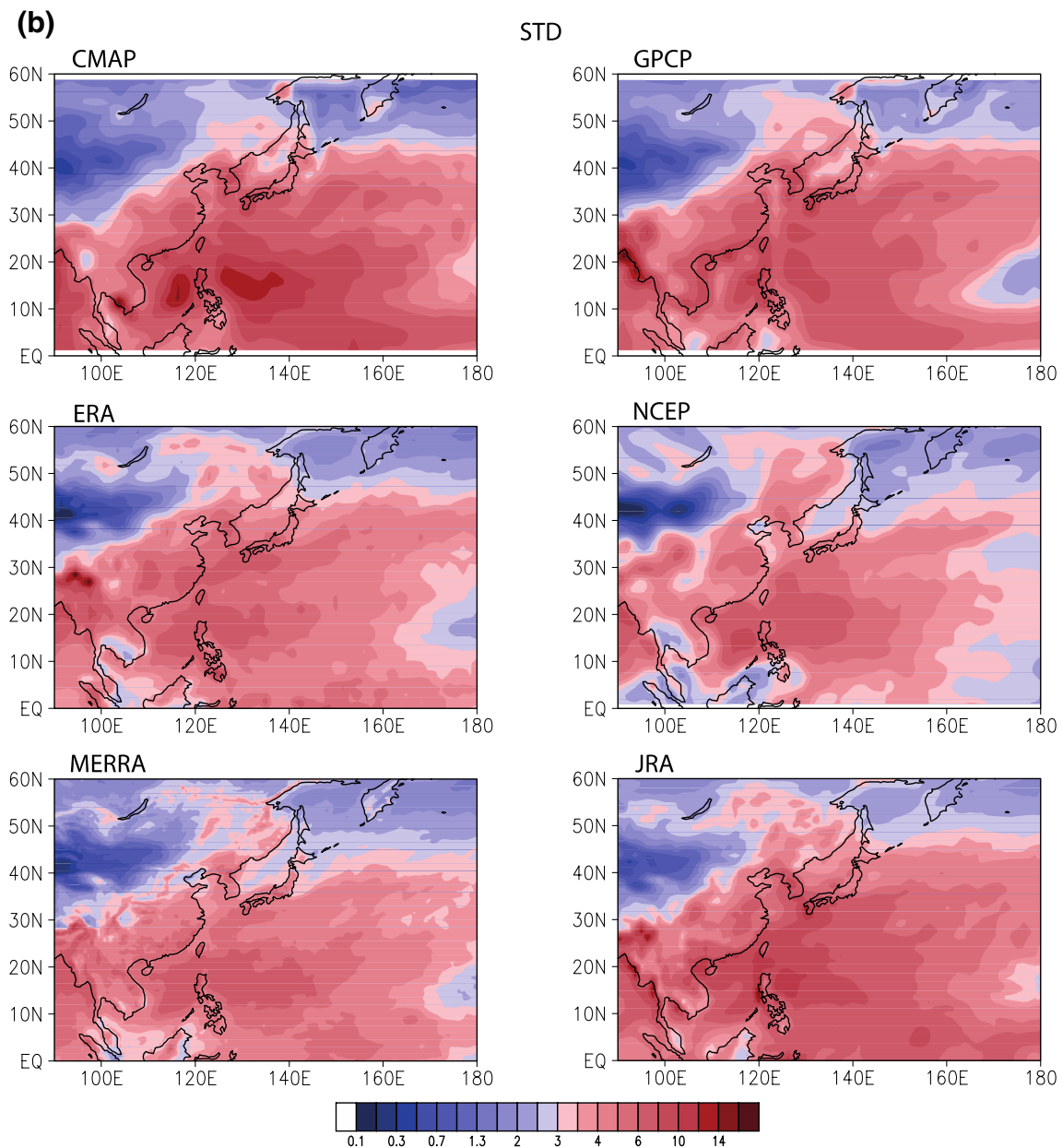


Fig. 1 (continued)

products except NCEP/NCAR overestimate the standard deviation over the continental region compared to CMAP; all of these reanalysis datasets except JRA underestimate the standard deviation over the ocean compared to CMAP. Over the tropical western Pacific, all reanalysis datasets except JRA show smaller means and standard deviations. The JRA reanalysis yields larger means but smaller standard deviations compared to CMAP. All the four reanalysis datasets also substantially underestimate the magnitude of the rain band over the ocean to the south of Japan, the most notable common problem in these reanalysis precipitation products. This problem is directly related to the

representation of the northward progression of the East Asian summer monsoon rainfall and is most (least) serious in the NCEP/NCAR (JRA) reanalysis. Overall, the spatial patterns of the long-term mean and the standard deviation of the reanalysis data agree well with CMAP when their performance is measured in terms of the spatial pattern correlation and the relative magnitude (Table 2). Table 2 shows that the reanalysis datasets yield larger pattern correlations with the CMAP data over ocean surfaces than over land surfaces; they also overestimate the magnitude of the spatial variability, especially for the mean climatology, compared to CMAP.

Table 2 A measure of difference in the means and standard deviations of the summer precipitation over the 37-year period in different precipitation datasets against the CMAP dataset in terms of the spatial pattern correlation (CORR) and the relative magnitude (MAG)

	GPCP	ECMWF	NCEP	MERRA	JRA
Mean					
CORR					
All	0.854	0.834	0.758	0.842	0.861
Ocean	0.908	0.904	0.837	0.899	0.884
Land	0.867	0.733	0.832	0.810	0.764
MAG					
All	0.927	1.033	0.965	0.983	1.203
Ocean	0.848	0.957	0.830	0.900	1.156
Land	1.201	1.299	1.442	1.277	1.366
Std					
CORR					
All	0.861	0.690	0.738	0.870	0.827
Ocean	0.877	0.781	0.777	0.918	0.801
Land	0.881	0.639	0.612	0.761	0.764
MAG					
All	0.905	0.787	0.716	0.750	0.982
Ocean	0.841	0.700	0.636	0.687	0.934
Land	1.145	1.116	1.020	0.991	1.163

The magnitude is the ratio of the spatial standard deviation of each dataset relative to the CMAP dataset

3.2 Comparison of individual CSEOF modes

In addition to the first two statistical moments, representations of various physical and/or dynamical processes involved in the spatiotemporal variations of summer precipitation over East Asia in each dataset are compared using the CSEOF components. For this evaluation, we first conduct CSEOF decomposition of each dataset. Then, regression analysis in CSEOF space is performed between the CMAP data (target) and the remaining five datasets (predictors). The R^2 values of regression for the first five CSEOF modes in Table 3 show that the GPCP data shows near-perfect regression fit for all of the first five modes. The four reanalysis

datasets also show close regression fits with the CMAP data for the first five CSEOF modes, although not as high as the GPCP dataset. This implies that the temporal evolution of the regressed CSLVs of GPCP and the four reanalysis datasets closely agree with those of CMAP.

The PC time series of the two observation-based datasets, CMAP and GPCP, agree closely with each other throughout the analysis period for the first three modes (Fig. 2). The first CSEOF mode represents the seasonal cycle, which dominates the evolution of summer monsoon rainfall over East Asia. The corresponding PC time series has a non-zero mean with strong fluctuations in the amplitude as shown in Fig. 2a for CMAP and GPCP. This implies that the phase of the evolution of the first mode as depicted in the corresponding loading vector (Supplemental Movie 1), remains positive throughout the analysis period. The PC time series (Fig. 2a) clearly shows the years of the monsoon precipitation well below (e.g., 1988, 1999, and 2000) or well above (e.g., 1990, 1992, and 2002) normal.

The temporal variations of the CSLVs of the first CSEOF mode show that not only GPCP, but also the four reanalysis datasets closely agree with CMAP throughout the 24 pentads (Supplemental Movie 1). That is, all the six datasets show similar evolutionary pattern of the summer monsoon. Similar evolution features can be obtained from a composite analysis over the 24 pentads (not shown) indicating that this is a repeating signal. Figure 3 measures the proximity, measured in terms of spatial pattern correlation (Fig. 3a) and relative RMSE (Fig. 3b) obtained by normalizing RMSE by the standard deviation of the CMAP data, of the first five CSLVs of the GPCP and the reanalysis datasets to those of CMAP for the 24 spatial patterns corresponding to the 24 pentads. The correlation coefficients and the relative RMSEs show that the CSLVs of GPCP are very close to those of CMAP. The pattern correlation between the CSLVs of GPCP and CMAP remain above 0.9 (black lines in Fig. 3a) with relative RMSEs generally below 0.4 (black lines in Fig. 3b) for the entire 24 pentads for the first five CSEOF modes. The R^2 values of regression for the GPCP data also exceed 0.98 for the first five modes (Table 3). The reanalysis datasets, except

Table 3 The R^2 values of regression for the five predictor variables (GPCP, ECMWF, NCEP, MERRA, JRA) against the CMAP pentad precipitation (target variable)

	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Total
CMAP	3.137	1.478	1.366	1.260	1.040	8.282
	12.06%	5.68%	5.25%	4.85%	4.00%	31.85%
GPCP	0.998	0.993	0.994	0.985	0.995	
ECMWF	0.993	0.991	0.993	0.976	0.977	
NCEP	0.991	0.978	0.975	0.975	0.964	
MERRA	0.993	0.990	0.978	0.976	0.960	
JRA	0.994	0.988	0.990	0.981	0.970	

The second and third rows represent eigenvalues and percent variance explained by the first five CSEOF modes

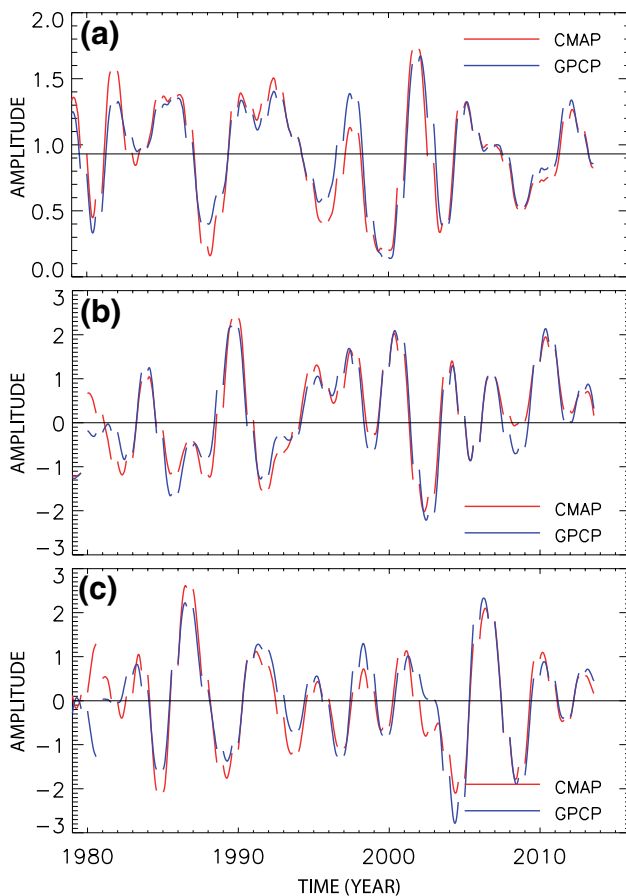


Fig. 2 The PC time series of the CMAP (red) and GPCP pentad precipitation (blue) for CSEOF modes **a** 1, **b** 2, and **c** 3

NCEP/NCAR, exhibit fairly high correlation with CMAP for the first five CSEOF modes; correlations averaged over the 24 pentads (colored thin lines in Fig. 3a) exceed 0.7. The NCEP/NCAR dataset yields slightly, but systematically, smaller correlation coefficients of around 0.6 compared to other three reanalysis datasets (0.7 or larger). The average relative RMSEs are below 0.7 for the reanalysis data except NCEP/NCAR for which relative RMSEs are around 0.8 for the first five CSEOF modes.

The second and the third CSEOF modes, which represent intraseasonal oscillations, of the five datasets also agree well with those of CMAP. Similar to the first CSEOF mode, the PC time series of CMAP and GPCP are close for the second and the third modes (Fig. 2b, c) with temporal correlations over 0.99 (Table 3). The temporal evolution of the loading vectors associated with the second and the third modes of the five datasets are close to those of the CMAP data over the entire 24 pentads (Fig. 3). Temporal evolutions of the second and third modes are shown in Supplemental Movies 2 and 3. Because of the irregularity in the phase of the intraseasonal oscillations, these two CSEOF modes are required to fully describe the intraseasonal oscillations of

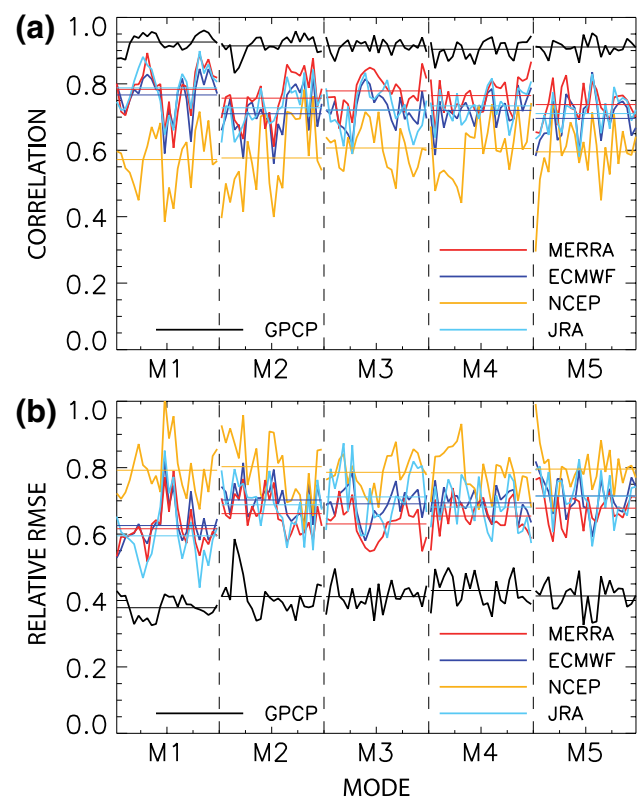


Fig. 3 **a** Pattern correlations and **b** relative RMSE of the first five CSEOF modes derived from the GPCP (black) and the four different reanalysis datasets (MERRA: red, ECMWF: blue, NCEP: yellow, and JRA: cyan). Each loading vector consists of 24 spatial patterns and there are 24 values of correlation and relative RMSE for each mode. The thin solid lines represent the correlation and relative RMSE averaged for each mode; there are five different colors for the five different datasets

precipitation. The four reanalysis datasets well reproduce the intraseasonal oscillations even though they slightly underestimate the magnitude of precipitation associated with these modes. This comparison of the PC time series and corresponding CSLVs shows that the two observation-based analysis datasets, CMAP and GPCP, agree closely with each other and that all the four reanalysis datasets perform reasonably in depicting the variability in the former datasets.

Figure 4 shows the Hovmöller diagram of the first CSEOF mode that correspond to the summer monsoon, averaged over the longitude range of 127.5°E–130°E. This longitude band is chosen in the context of the location of the East Asian monsoon. The monsoon rain band arrives 33°N, around the latitude of the southern-most territory of Korea, on around July 20. The monsoon rainfall arrives in its northern-most reach of 38°N in early August in a climatological sense. The GPCP data agree closely with the CMAP data in the northward march of the summer rainfall even though the magnitude of the anomalous rainfall is slightly underestimated compared to CMAP, especially to the south

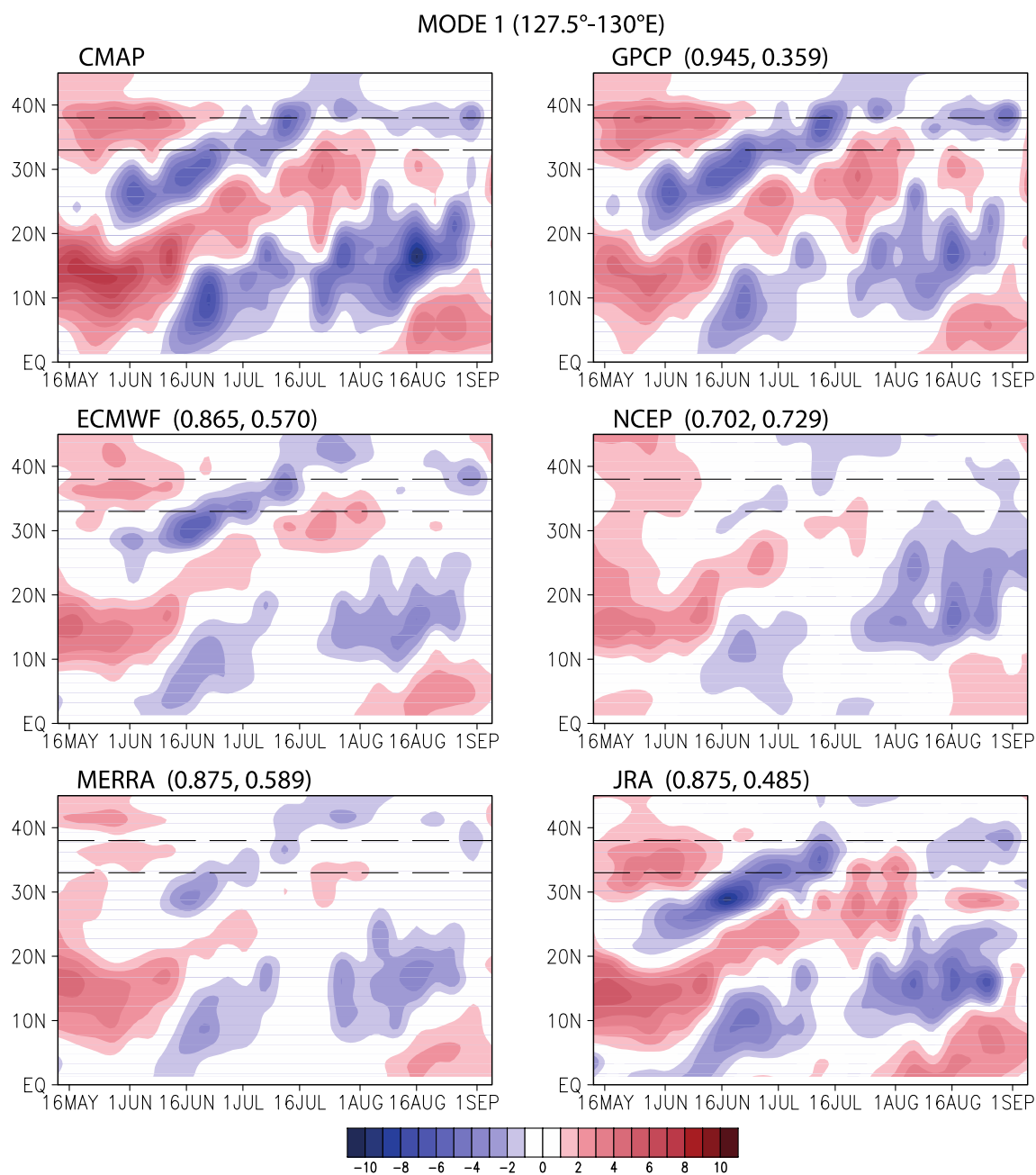


Fig. 4 The Hovmöller diagram of the longitude (127.5°–130°E) averaged seasonal cycle (summer monsoon) derived from the six datasets for the CSEOF mode 1. The two dashed lines show the latitude range

of 30°N. The JRA and ERA-Interim reanalysis datasets reasonably depict the northward movement of the monsoon rainfall in the region. The NCEP and MERRA reanalysis datasets, on the other hand, seriously underestimate the magnitude of the anomalous precipitation within the longitude range, especially in the southern half of the Korean Peninsula (33°N–38°N, marked by the dashed lines in Fig. 4).

Figures 5 and 6 show the Hovmöller diagram of the second and the third CSEOF modes averaged over the longitude

of the Republic of Korea. The two numbers in parenthesis denote the correlation coefficient and the relative RMSE with respect to the CMAP data

range of 100°E–140°E which encompasses the East Asian monsoon region. Similar to the first CSEOF mode shown in Fig. 4, CMAP and GPCP agree closely with each other; however, magnitudes of the anomalous precipitation in the GPCP data are generally weaker than in the CMAP data. All the six datasets capture the overall patterns of the intraseasonal oscillation during summer. The timing and the sign of the intraseasonal oscillations in the reanalysis datasets agree reasonably with the two observation-based datasets,

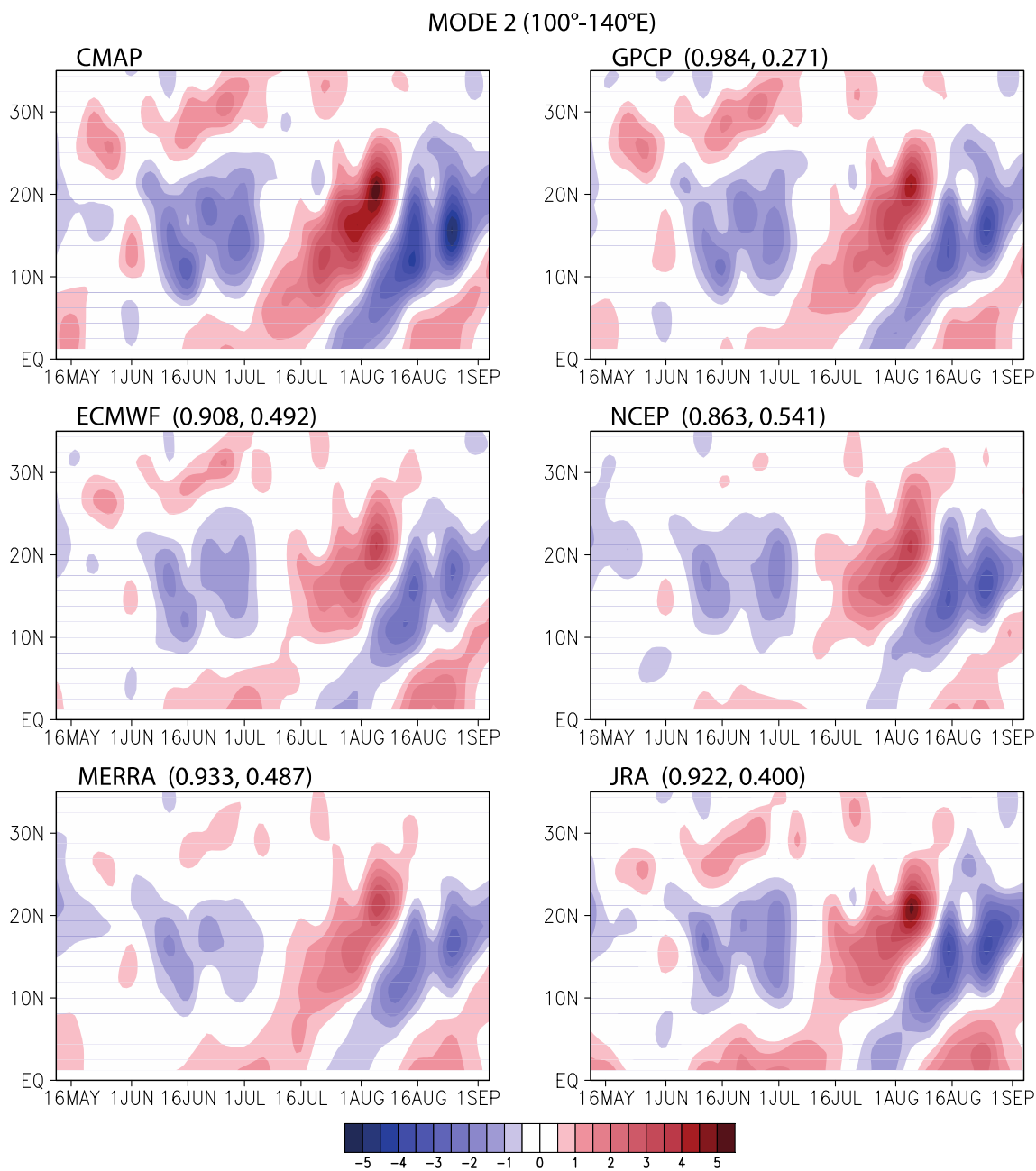


Fig. 5 Same as Fig. 4, but for the CSEOF mode 2

specifically in the middle and late summer, July and August. The precipitation anomalies of MERRA and JRA agree closely with those of the two observation-based datasets for July and August over the entire latitude range from the Equator to 35°N. The NCEP/NCAR reanalysis shows problems in depicting the precipitation anomalies corresponding to intraseasonal oscillations in the tropical regions between the Equator and 10°N. Despite the general agreement between the reanalysis datasets and the observation-based CMAP and GPCP datasets, all the four reanalysis datasets underestimate, compared to the CMAP/GPCP data, the precipitation

variability associated with intraseasonal oscillations, especially in the tropics for June.

Figure 7 shows the correlation coefficients over the 24 pentads of the CSLVs between the reanalysis datasets and CMAP for the first three CSEOF modes. Unlike conventional EOF analysis where the loading vector is uniform for all time steps, each time step (pentad in this study) of the nested period has its own spatial pattern in CSEOF analysis. The correlation coefficient for the first mode generally exceeds 0.5 over most of the East Asia domain for all of the reanalysis datasets except NCEP (Fig. 7a). All the four

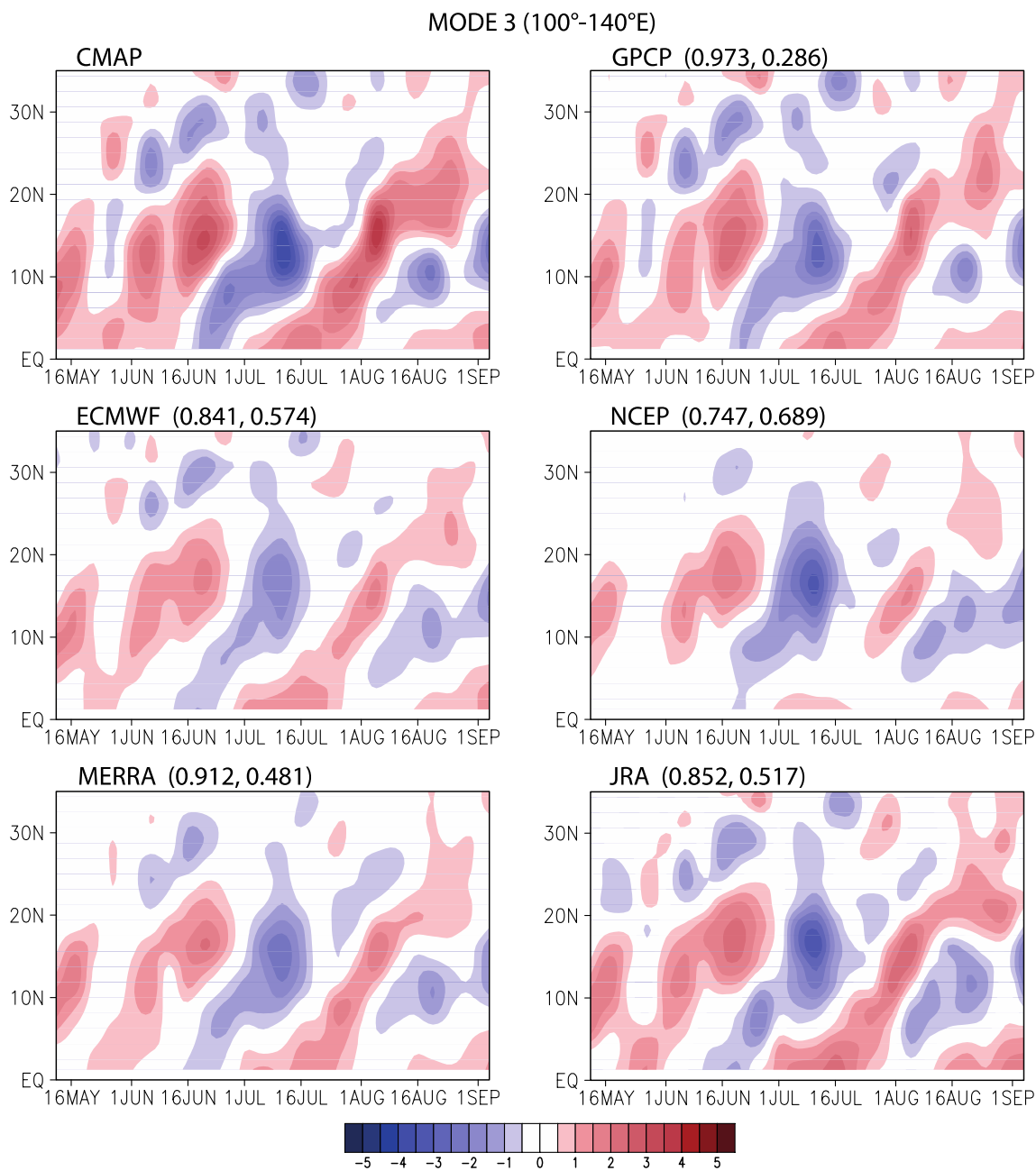
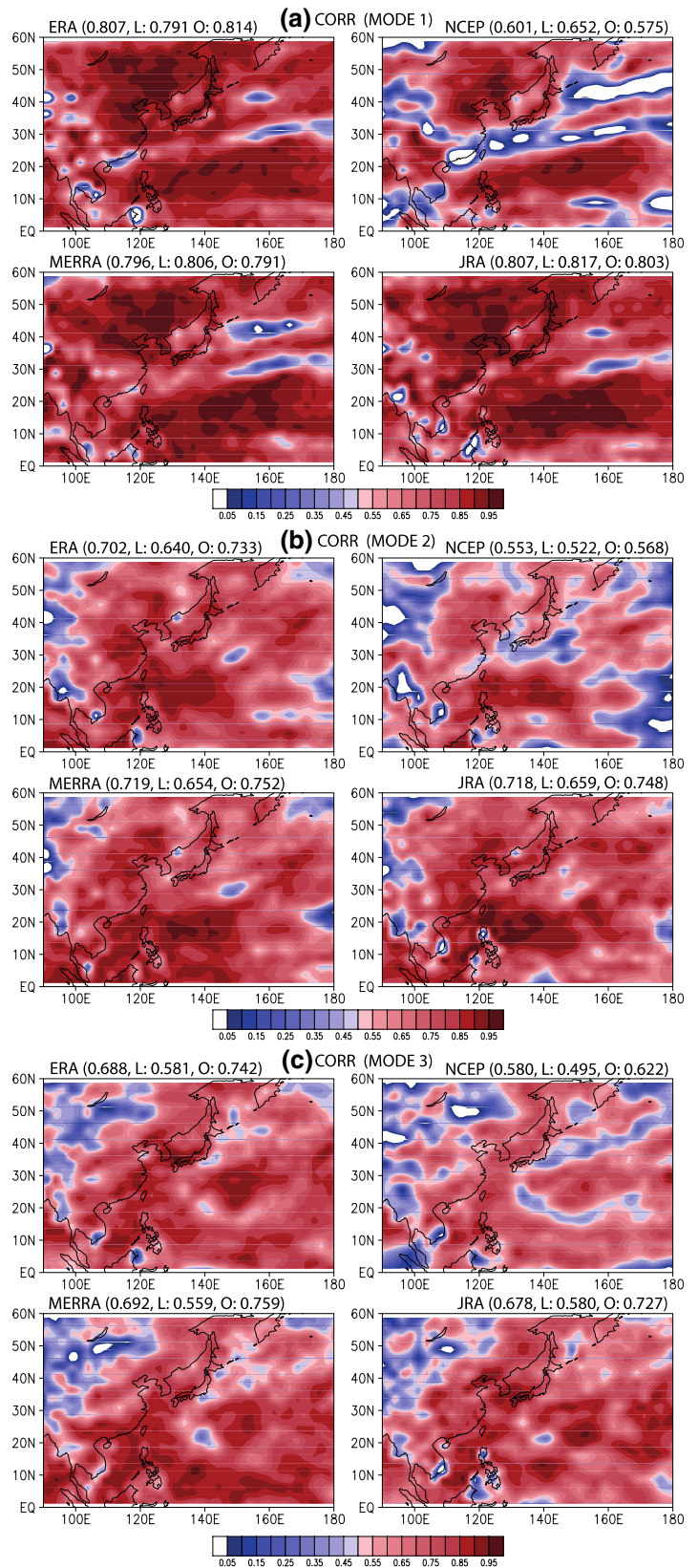


Fig. 6 Same as Fig. 4, but for the CSEOF mode 3

reanalysis datasets, to some degree, show common weaknesses in representing the summer monsoon rainband over the northwestern Pacific to the south of Japan, especially to the east of 150°E. This indicates discrepancy in the timing and/or the position of the rainband in the reanalysis datasets relative to the CMAP data. In general, correlations between CMAP and the four reanalysis datasets are fairly high in the mid-latitude continental regions and over the subtropical western Pacific. All of the reanalysis datasets also deviate notably from CMAP for some isolated regions, mostly in mountainous and/or coastal regions. This may be due to

the reanalysis models' inability to represent the effects of complex orography and/or sparse observational inputs for assimilations. It must be noted that the observed CMAP and GPCP data also suffer from the lack of gauge data in these regions (e.g., Kim et al. 2015a; Kim and Park 2016). For higher modes (Fig. 7b, c), correlations decrease to below 0.5 over much of the inland regions of the Asian continent. Along the continental margin, correlation coefficients are still reasonable except for the Indochina Peninsula, Myanmar and Malaysia. For oceanic regions, all of the reanalysis products exhibit small correlation coefficients over the

Fig. 7 The temporal correlation for the first three CSEOF loading vectors derived from the pentad precipitation of four different reanalysis datasets against those of the CMAP pentad precipitation. The numbers in parenthesis represent domain average, land average, and ocean average values



region of the rainband to the south of Japan and over the subtropical central Pacific Ocean. Performance of the reanalysis models for reproducing loading vectors over the continental interior regions, specifically the Tibetan Plateau and the Inner Mongolia, generally deteriorates for higher modes. The pattern correlation between the CSLVs of CMAP and GPCP is fairly high over much of the analysis domain for all three modes (figure not shown) with the average pattern correlation being over 0.9 (see Fig. 3a).

Figure 8 shows the relative RMSE of CSLVs calculated over the 24 pentads between the reanalysis datasets and the CMAP data for the first three modes. The relative RMSE for the first mode is generally well below unity except over the interior regions of East Asia, Myanmar and eastern India (Fig. 8a). Regions of large relative RMSE over the western Pacific Ocean generally agree with the regions of small correlation (Fig. 7a). Thus, most notable errors in these reanalysis datasets are associated with the representation of the main (monsoon) rainband over the mid-latitude western Pacific during summer. Similar to the first mode, the second and third modes (Fig. 8b, c) exhibit larger relative RMSEs over the inland area and the high-latitude regions of the Pacific Ocean compared to other East Asia regions within the domain. Like the first mode, the regions of large relative RMSE generally coincide with the regions of small correlations, and the relative RMSE tends to be larger for higher modes. That is, weakness in representing the processes associated with the first three modes appears in the same geographical regions. This suggests common problems such as the lack of input observations and/or difficulties in simulating the effects of complex orography in these reanalysis models.

3.3 Reconstructed CSEOF modes

Summer precipitation over East Asia represented in the first five CSEOF modes are reconstructed as:

$$\{T^{(N)}(r, t), P^{(N)}(r, t)\} = \sum_{n=1}^N \{B_n(r, t), C_n^{(reg)}(r, t)\} T_n(t), \quad (8)$$

where N denotes the number of modes retained to reconstruct the precipitation field; $N = 5$ since the first five CSEOF modes are used. Note that the regressed loading vectors are used in conjunction with the target PC time series to compare the precipitation variability between the CMAP (target) and the four reanalysis datasets (predictors) associated with the same dynamical processes. Note that the reconstructed GPCP data (not shown) are nearly identical to the reconstructed CMAP data and are not included in this evaluation. The reconstruction based on the first three CSEOF modes yields nearly the same results as the five-mode reconstruction presented below except that the three-mode

reconstruction yields slightly smaller error standard deviations for both land and ocean surfaces (not shown).

Figure 9 shows the correlation (Fig. 9a), the relative RMSE (Fig. 9b) and the error standard deviation (Fig. 9c) between the 5-mode reconstructed CMAP data (i.e., target) and the four reanalysis datasets (i.e., predictors). Correlations between the reconstructed CMAP and the reconstructed reanalysis datasets are reasonably high in the entire domain except for the inland region of East Asia and Myanmar (Fig. 9a). The relative RMSE is also reasonably small except for the interior region of the continental East Asia. Error standard deviations are ~ 2 – 3 times the magnitude of the standard deviation of the five-mode reconstructed CMAP data in the region to the south of 40°N , both over the land and ocean. While correlations are fairly large and relative RMSEs are fairly small over the subtropical western Pacific, error standard deviation is large, especially over the subtropical Pacific Ocean. This indicates that the reanalysis errors in representing the precipitation variability associated with the processes represented in the first five modes tend to increase as the variability (and the mean) of the CMAP data increases. The region of high error standard deviations includes the major precipitation regions. In the high latitudes to the north of 40°N , error standard deviation decreases notably, indicating that errors also decrease as the precipitation intensity decreases.

3.4 Variability not included in the first five modes

The remaining precipitation variability not included in the first five CSEOF modes is defined by

$$\{T^{(r)}(r, t), P^{(r)}(r, t)\} = \{T(r, t) - T^{(5)}(r, t), P(r, t) - P^{(5)}(r, t)\}. \quad (9)$$

The superscript r in (9) denotes the remaining variability. Figure 10 shows the correlation, the relative RMSE and the error standard deviation over the 24 pentads between the reanalysis datasets and the CMAP data after removing the first five CSEOF modes. All the four reanalysis datasets show correlations with CMAP below 0.5 over the Tibetan Plateau and the Indochina Peninsula, the Philippines, and the Maritime Continent (Fig. 10a). All of the reanalysis datasets except MERRA also show smaller correlations over the tropical western Pacific. Other than these regions, the four reanalysis datasets show good correlations with the CMAP data, especially over the mid-latitude region (Fig. 10a). The relative RMSE with respect to the CMAP data also show similar regional variations as the correlation coefficients. The geographical variations of the relative RMSE (Fig. 10b) corresponds inversely to the correlation coefficient (Fig. 10a); i.e., larger (smaller) relative RMSEs over the regions of smaller (larger) correlation coefficients. Visual inspections show that the MERRA (NCEP/NCAR) data

Fig. 8 The relative RMSE for the first three CSEOF loading vectors derived from the pentad precipitation of four different reanalysis datasets against those of the CMAP pentad precipitation. The numbers in parenthesis represent domain average, land average, and ocean average values

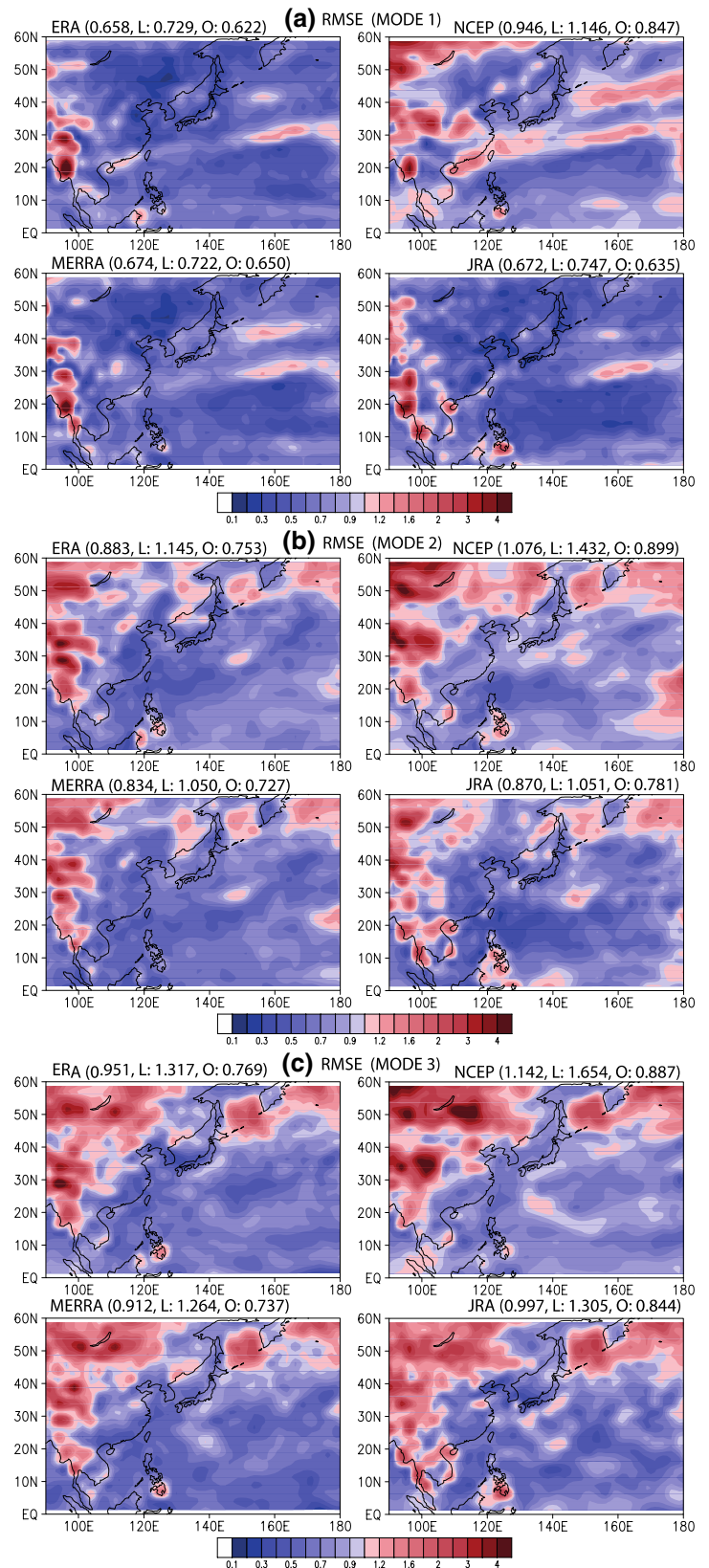


Fig. 9 The correlation coefficients (upper), relative RMSEs (middle), and error standard deviations (lower) of the reconstructed precipitation based on the first five CSEOF modes derived from the four reanalysis products against that of the CMAP pentad precipitation. The first five CSEOF modes explain ~32% of the total variability of the CMAP precipitation. The numbers in parenthesis represent domain average, land average, and ocean average values

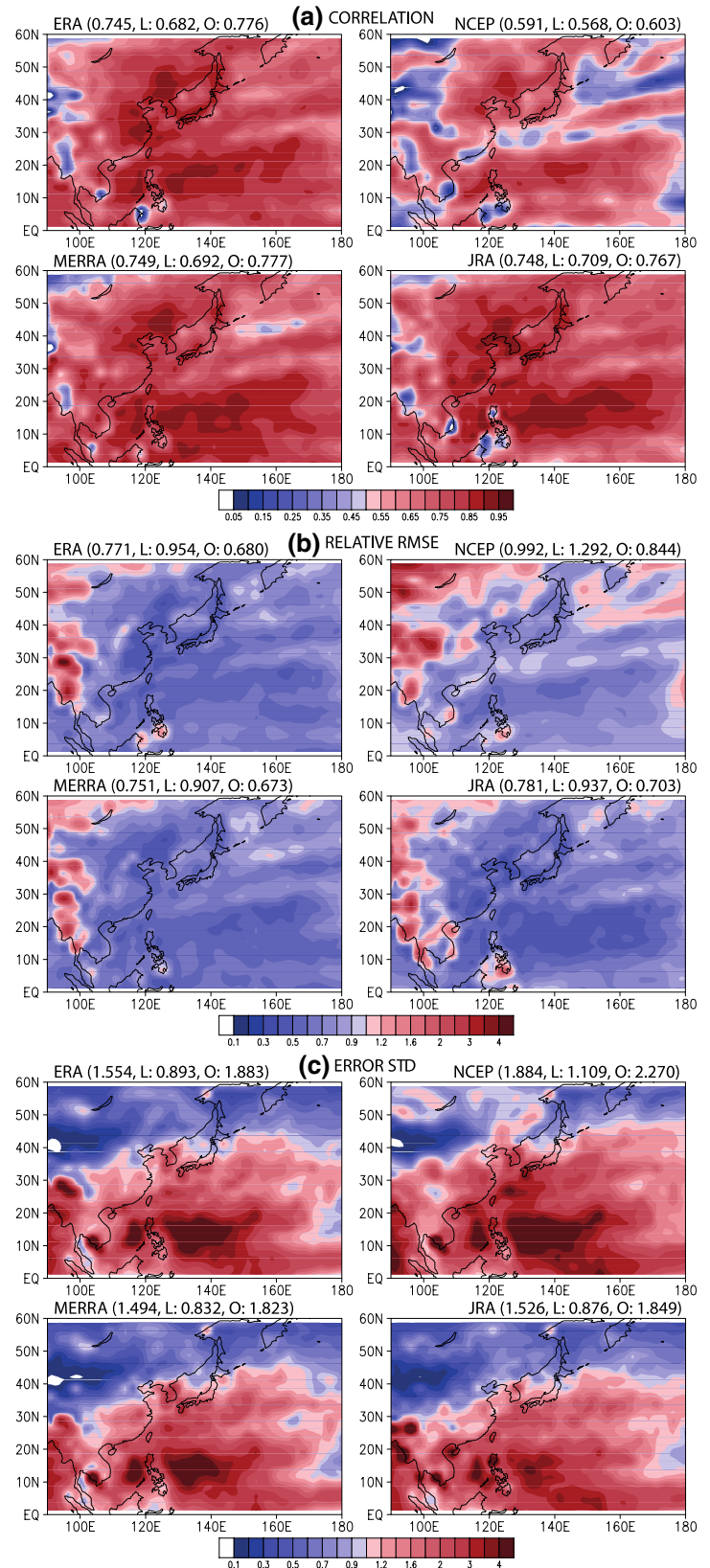
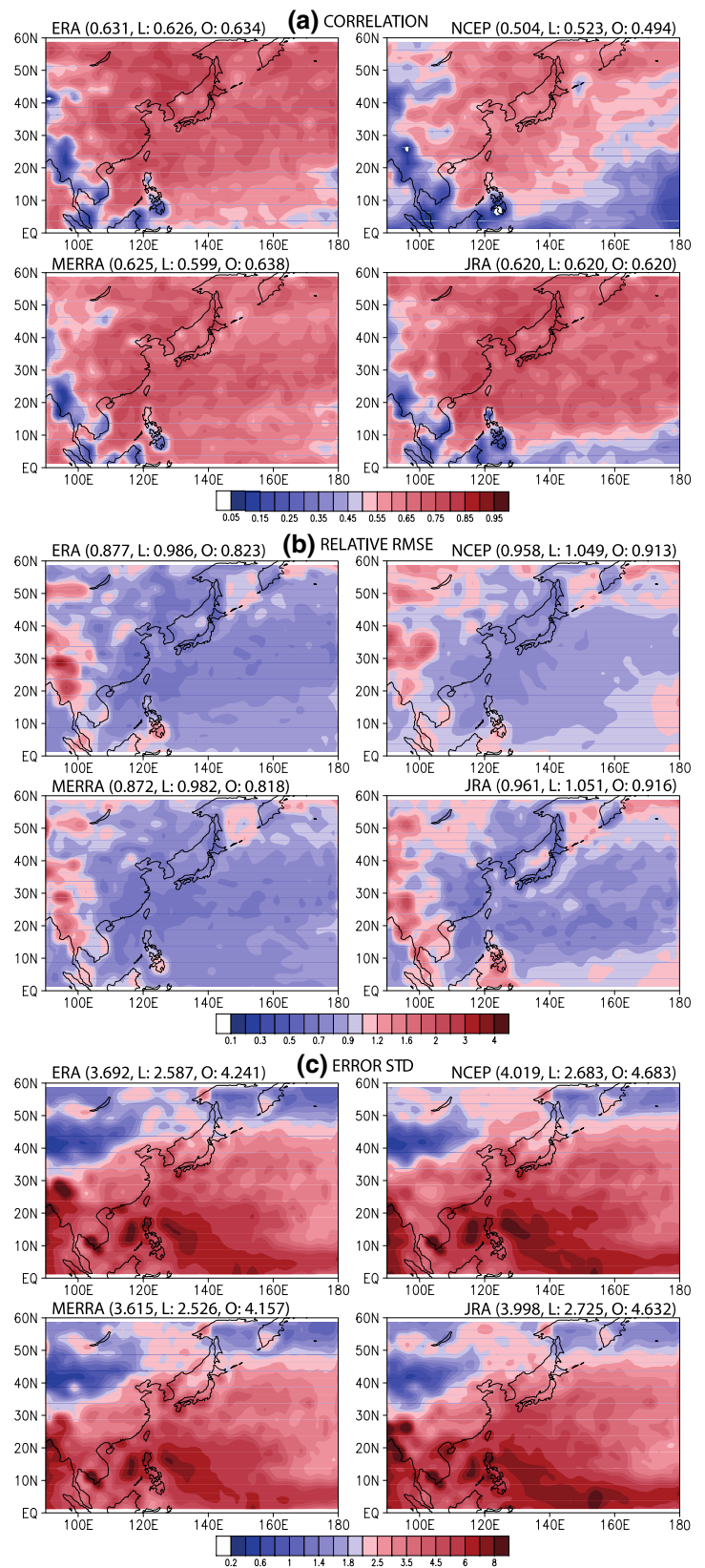


Fig. 10 The correlation coefficients (upper), relative RMSEs (middle), and error standard deviations (lower) of the remaining precipitation variability between the four reanalysis products and the CMAP pentad precipitation. The remaining precipitation variability is obtained by removing the five-mode reconstruction from the raw data. The remaining precipitation variability explains ~68% of the total variability of the CMAP precipitation. The numbers in parenthesis represent domain average, land average, and ocean average values



yields the largest (smallest) area of the correlation > 0.5 and the relative RMSE < 1.0 in the East Asian domain analyzed in this study. Error standard deviation (Fig. 10c) exceeds 5 mm day^{-1} in the lower-latitude regions below 30°N ; it is generally larger than that for the reconstructions using the first five CSEOF modes in Fig. 9c (note the scale differences between Figs. 9c, 10c). Considering that the remaining modes represent about 68% of the total summer precipitation variability in the domain, the four reanalysis datasets possess acceptable accuracy in representing summer precipitation characteristics over much of East Asia and the western Pacific Ocean, especially in the mid-latitude land and ocean areas. The regions of common weakness include the tropical western Pacific Ocean and the southern part of the East Asia land surfaces, especially the region including the southern and eastern slope of the Tibetan Plateau.

4 Discussions and conclusions

This study has evaluated the observation-based GPCP and four reanalysis precipitation datasets against the observation-based CMAP data for the variability associated with each CSEOF mode as well as for the first two statistical moments. Based on CSEOF analysis, precipitation in these six datasets has been decomposed into uncorrelated (nearly independent) and orthogonal evolution patterns. Then, one-to-one correspondence between the CSEOF modes derived from the target dataset (CMAP) and the predictor datasets (GPCP, ERA-Interim, NCEP, MERRA, and JRA) is established using the regression analysis in CSEOF space. The resulting regressed CSEOF modes as well as the reconstructed data based on them are compared against the corresponding fields of the CMAP data in order to assess the closeness between the predictor datasets and the CMAP data, i.e., to evaluate the predictor datasets against the target (CMAP) data. The analysis period includes 120 days (24 pentads) for each of the 37 summers from 1979 to 2015. It should be noted that using GPCP as the target variable yields essentially the same results obtained in this study.

Evaluations of the first two moment statistics, individual CSEOF modes, and the reconstructions for various combinations of the CSEOF modes, indicate that the GPCP data agrees closely with the CMAP data, i.e., the two observation-based precipitation analysis datasets represent essentially identical summer precipitation characteristics over East Asia and the western Pacific Ocean. The closeness of the GPCP and CMAP datasets are also supported from other metrics such as the spatial and temporal correlations, relative RMSEs, and error standard deviations calculated between these two datasets. The close agreement between CMAP and GPCP, especially over the land surface, is not surprising because these two datasets are based on similar

input data and are calibrated using surface gauge data. The four reanalysis datasets also agree well with the CMAP data. The most notable problems in the reanalysis datasets are the underestimation of the summer rainband over the subtropical western Pacific to the south of Japan and in the tropical western Pacific Ocean. These errors are smallest in the JRA and the ERA-Interim datasets.

Assessment of the accuracy of the four reanalysis precipitation data in terms of the CSEOF loading vectors and PC time series reveals much more than what conventional methods can do. The reanalysis models are reasonably accurate in reproducing the first three modes of variability including the seasonal cycle (summer monsoon) and the intraseasonal oscillations (the second and third modes). Evaluations of the first three CSEOF modes from the four analysis datasets also show that they are fairly reasonable in terms of reproducing the evolution patterns of summer precipitation over East Asia and the western Pacific associated with these modes in CMAP. The most notable shortcomings of the reanalysis datasets are the underestimation of precipitation over the mid-latitude western Pacific to the south of Japan, which shows up clearly in the mean summer precipitation (Fig. 1). Figure 7a shows that this problem is mostly due to the lack of representation of the summer monsoon rainfall associated with the first EOF mode.

The first five CSEOF modes of the four reanalysis datasets are fairly similar to those derived from the CMAP data with pattern correlations above 0.7 and average relative RMSEs below 0.7 except for NCEP/NCAR. The relatively poor skill of the NCEP/NCAR datasets compared to the other three datasets may be related with the fact that it is the earliest products among the four; the NCEP/NCAR reanalysis was generated by the earliest generation model and datasets without the benefit of the recently developed model formulations and observational datasets.

The reconstructed precipitation using the first five CSEOF modes for the four reanalysis datasets compares faithfully with that for the CMAP data. Correlations (RMSEs) between the reconstructed CMAP and reanalysis data are reasonably large (small) in the entire East Asian domain except for the inland region of East Asia and Myanmar. Error standard deviation is large in the region to the south of 40°N where precipitation magnitude is also large. The remaining precipitation outside the first five CSEOF modes agrees reasonably with the corresponding precipitation in the CMAP data as well. Both correlation coefficients and relative RMSEs deteriorate somewhat compared to those from the first five modes. Error standard deviation of the remaining precipitation is larger than that for the first five CSEOF modes. Despite the shortcomings discussed above, the reconstructed reanalysis data reasonably agree with the corresponding CMAP data in most of the East Asian region and the western Pacific Ocean. All the reconstructed reanalysis data for

the first five CSEOF modes as well as the remaining precipitation outside the first five modes, show relatively poor performance over the subtropical western Pacific, the eastern Indian Ocean off the west coast of the Indochina Peninsula, and the land areas including the Indochina Peninsula, the Maritime Continent, and the southern/eastern slope of the Tibetan Plateau compared to other regions.

All of the reanalysis products examined in this study need improvement in simulating precipitation and associated processes over the western Pacific. Evaluation in this study shows that reanalysis models generally do not reproduce well the evolution of summer rainband over the ocean, especially over the mid-latitude eastern Pacific to the south of Japan and its subsequent movement over eastern China, Korea, and Japan. This is a critical concern because these regions contain massive populations and industrial activities that are heavily affected by summer rainfall. Problems over the land surfaces of highly complex terrain and/or coastal geometries, at least partially, may be related with the difficulties in representing orography in numerical models used for reanalysis. These regions are also poorly instrumented, i.e., inputs for reanalysis is much scarcer than other regions of higher observation density; this is a problem not only in assimilating precipitation for reanalysis but also in producing observed precipitation analysis (e.g., Kim et al. 2015a).

Acknowledgements The CMAP, NCEP/NCAR, GPCP, MERRA, ERA-Interim, and JRA reanalysis datasets are obtained from ESRL/NOAA, ESRL/NOAA, GSFC/NASA, GSFC/NASA, ECMWF, and JMA, respectively. This study is supported by the National Institute of Meteorological Sciences (NIMS)–Korean Meteorological Administration (NIMS-2016-3100). KYK is supported by the National Science Foundation of Korea under the Grant number NRF-2017R1A2B4003930.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adler RF, Huffman GJ, Chang A, Ferraro R, Xie P, Janowiak J, Rudolf B, Schneider U, Curtis S, Bolvin D, Gruber A, Susskind J, Arkin P, Nelkin E (2003) The version 2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *J Hydrometeorol* 4:1147–1167
- Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Kohler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J, Park B, Peubey C, de Rosnay P, Tovelato C, Thepaut J, Vitart F (2011)

- The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q J R Meteorol Soc* 137:553–597
- Guirguis KJ, Avissar R (2008) An analysis of precipitation variability, persistence, and observational data uncertainties in the western United States. *J Hydrometeorol* 9:843–865
- Harada Y, Kamahori H, Kobayashi C, Endo H, Kobayashi S, Ota Y, Onoda H, Onogi K, Miyaoka K, Takahashi K (2016) The JRA-55 reanalysis: representation of atmospheric circulation and climate variability. *J Meteorol Soc Jpn* 94:262–302. <https://doi.org/10.2151/jmsj.2016-015>
- Huffman GJ, Adler RF, Arkin P, Chang A, Ferraro R, Gruber A, Janowiak J, McNab A, Rudolf B, Schneider U (1997) The global precipitation climatology project (GPCP) combined data set. *Bull Am Meteorol Soc* 78:5–20
- Huffman GJ, Adler RF, Bolvin DT, Gu G (2009) Improving the global precipitation record: GPCP version 2.1. *Geophys Res Lett* 36:L17808. <https://doi.org/10.1029/2009GL040000>
- Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc* 77:437–470
- Kim K-Y, North GR (1997) EOFs of harmonizable cyclostationary processes. *J Atmos Sci* 54:2416–2427
- Kim J, Park SK (2016) Uncertainties in calculating precipitation climatology in East Asia. *Hydrol Earth Syst Sci* 20:651–658. <https://doi.org/10.5194/hess-20-651-2016>
- Kim K-Y, North GR, Huang J (1996) EOFs of one-dimensional cyclostationary time series: computations, examples and stochastic modeling. *J Atmos Sci* 53:1007–1017
- Kim J, Waliser D, Mattmann C, Goodale C, Hart A, Zimdars P, Crichton D, Jones C, Nikulin G, Hewitson B, Jack C, Lennard C, Favre A (2013) Evaluation of the CORDEX-Africa multi-RCM hindcast: systematic model errors. *Clim Dyn*. <https://doi.org/10.1007/s00382-013-1751-7>
- Kim J, Sanjay J, Mattmann C, Boustani M, Ramarao MVS, Krishnan R, Waliser D (2015a) Uncertainties in estimating spatial and interannual variations in precipitation climatology in the India–Tibet region from multiple gridded precipitation datasets. *Int J Climatol*. <https://doi.org/10.1002/joc.4306>
- Kim K-Y, Hamlington BD, Na H (2015b) Theoretical foundation of cyclostationary EOF analysis for geophysical and climatic variables: concepts and examples. *Earth Sci Rev* 150:201–218. <https://doi.org/10.1016/j.earscirev.2015.06.003>
- Kim J, Waliser DE, Cesana G, Jiang X, L’Ecuyer T, Neena JM (2018) Cloud and radiative heating profiles associated with the boreal summer intraseasonal oscillations. *Clim Dyn* 50:1485–1494. <https://doi.org/10.1007/s00382-017-3700-3>
- Kobayashi S, Coauthors (2015) The JRA-55 Reanalysis: General specifications and basic characteristics. *J Meteorol Soc Jpn* 93:5–48. <https://doi.org/10.2151/jmsj.2015-001>
- Legates D, Willmott C (1990) Mean seasonal and spatial variability in gauge-corrected, global precipitation. *Int J Climatol* 10:111–127
- Neena JM, Waliser DE, Jiang X (2016) Model performance metrics and process diagnostics for boreal summer intraseasonal variability. *Clim Dyn* 48:1661–1683
- Prakash S, Mitra AK, Momin IM, Rajagopal EN, Basu S, Collins M, Turner AG, Rao KA, Ashok K (2014) Seasonal intercomparison of observational rainfall datasets over India during the southwest monsoon season. *Int J Climatol*. <https://doi.org/10.1002/joc.4129>
- Preethi B, Mujumdar M, Kripalani R, Prabhu A, Krishnan R (2017) Recent trends and teleconnections among South and East Asian monsoons in a warming environment. *Clim Dyn* 48:2489–2505

- Rienecker MM et al (2011) MERRA: NASA's modern-era retrospective analysis for research and applications. *J Clim* 24:3624–3648
- Vogel R (2013) Quantifying the uncertainty of spatial precipitation analyses with radar-gauge observation ensemble. Scientific Report MeteoSwiss No. 95, Federal Office of Meteorology and Climatology, MeteoSwiss, Zürich, Switzerland. <http://www.meteoschweiz.ch>
- Whitehall K, Mattmann C, Waliser DE, Kim J, Goodale C, Hart A, Ramirez P, Zimdars P, Crichton D, Jenkins G, Jones C, Asrar G, Hewitson B (2012) Building model evaluation and decision support capacity for CORDEX. *WMO Bull* 61:29–34
- Xie P, Arkin PA (1995) An intercomparison of gauge observations and satellite estimates of monthly precipitation. *J Appl Meteorol* 34:1143–1160
- Xie P, Arkin PA (1997) Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull Am Meteorol Soc* 78:2539–2558
- Yatagai A, Kamiguchi K, Arakawa O, Hamada A, Yasutomi N, Kitoh A (2012) APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *B Am Meteorol Soc* 93:1401–1415