A Moving Updated Statistical Prediction Model for Summer Rainfall in the Middle-Lower Reaches of the Yangtze River Valley

YAN GUO

State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing, and Zhuhai Joint Innovative Center for Climate–Environment–Ecosystem, Future Earth Research Institute, Beijing Normal University, Zhuhai, China

JIANPING LI[a]

State Key Laboratory of Earth Surface Processes and Resource Ecology, and College of Global Change and Earth System Science, Beijing Normal University, and Joint Center for Global Change Studies, Beijing, China

JIANGSHAN ZHU

Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

(Manuscript received 22 November 2016, in final form 25 April 2017)

ABSTRACT

Because summer rainfall in the middle-lower reaches of the Yangtze River valley has remarkable interannual and decadal variability and because the precursors that modulate the interannual rainfall change with the decadal variation of the background state, a new model that employs a novel statistical idea is needed to yield an accurate prediction. In this study, the interannual rainfall model (IAM) and the decadal rainfall model (DM) were constructed. Moving updating of the IAM with the latest data within an optimal length of training period (20 yr) can partially offset the effect of decadal change of precursors in IAM. To predict the interannual rainfall of 2001–13 for validation, 13 regression models were fitted with precursors that change every 4–5 yr, from the preceding winter North Atlantic Ocean sea surface temperature anomaly (SSTA) dipole to the Mascarene high, followed by the East Asia sea level pressure anomaly (SLPA) dipole and the preceding autumn North Pacific SSTA dipole. The moving updated model demonstrated high skill in predicting interannual rainfall, with a correlation coefficient of 0.76 and a hit rate of 76.9%. The DM was linked to the April SLPA in the central tropical Pacific Ocean, and it maintained good performance in the testing period, with a correlation coefficient of 0.77 and a root-mean-square error (RMSE) of 7.7%. The statistical model exhibited superior capability even when compared with the best forecast by the Climate Forecast System, version 2 (CFSv2), initiated in early June, as indicated by increased correlation coefficient from 0.62 to 0.75 and reduced RMSE from 12.3% to 10.7%.

1. Introduction

Seasonal rainfall prediction is of great importance to the survival and development of humanity because it is in high demand for agriculture, water resource management, and the energy and transportation sectors. As a part of the East Asian summer monsoon (EASM) major rainfall belt, mei-yu–baiu–changma, summer rainfall over the middle-lower reaches of the Yangtze River valley (YRV) exhibits large interannual and decadal variability, and the induced droughts and floods can cause severe economic loss and casualties, as occurred from flooding in 1998. Understanding the mechanism of YRV summer rainfall variation and providing steadily reliable predictions are crucial to national disaster prevention and mitigation.

Currently, seasonal prediction with fully coupled climate models has become routine in a number of national climate centers worldwide. However, rainfall prediction skill with climate models remains limited, especially for the EASM rainfall (Lee et al. 2011).

[a] Additional affiliation: Laboratory for Regional Oceanography and Numerical Modeling, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China.

Corresponding author: Dr. Jianping Li, ljp@bnu.edu.cn

DOI: 10.1175/JAMC-D-16-0376.1

A statistical prediction model is an alternative to enhance prediction skill in such a situation. As in-depth studies have been performed to understand the mechanism of YRV summer rainfall variation, sea–land–atmosphere precursors that influence rainfall variation have increasingly been uncovered, such as the decaying and developing phases of El Niño–Southern Oscillation (ENSO) (Xie et al. 2016; Wang et al. 2000; Wang et al. 2009), spring North Atlantic Oscillation/Antarctic Oscillation (NAO/AO) (Gong et al. 2011; Wu et al. 2009), preceding winter Arctic sea ice (Wu et al. 2004), preceding winter snow depth over the Tibetan Plateau (Wei et al. 1998; Zhu et al. 2009), and atmospheric circulations in the Southern Hemisphere (Fan 2006; Liu et al. 2008; Nan and Li 2003; Xue et al. 2003). With these precursors, statistical models have been developed to predict summer rainfall over the YRV. Fan et al. (2008) proposed a new approach to predict YRV summer rainfall by predicting the year-to-year increment of rainfall using a multilinear regression equation that consists of six precursors, including AO and the meridional wind shear between 850 and 200 hPa over the Indo-Australian region. By combining ENSO and spring NAO, Wu et al. (2009) successfully predicted the EASM strength that is accurately represented by YRV summer rainfall. Wu and Yu (2016) constructed a partial least squares (PLS) model to predict the EASM strength using two leading PLS modes associated with mega-ENSO.

Along with EASM decadal weakening, summer rainfall over the YRV has intensified since the 1970s, followed by a continuing reduction after 2000. The decadal variation of the YRV summer rainfall has attracted considerable attention. Ping et al. (2006) and Wei (2006) indicated that there are different sea–land–atmosphere precursors modulating the YRV summer rainfall variations on interannual and decadal time scales. Therefore, it is necessary to differentiate the interannual and decadal variations when we attempt to identify precursors and develop statistical prediction models, which means that the interannual rainfall model (IAM) and the decadal rainfall model (DM) are constructed, respectively. Time-scale decomposition has been demonstrated as an effective approach to statistically predict summer rainfall over northern China, which also has interannual and decadal variations that are modulated by different sea–land–atmosphere precursors (Guo et al. 2012; Ruan and Li 2016).

The sea–land–atmosphere precursors that modulate the interannual rainfall variation are not immutable. They might change with the decadal variation of the background state. Many studies have focused on the decadal change of the relationship between interannual EASM rainfall and the related sea–land–atmosphere precursors. It was indicated that the relationship between ENSO and EASM-related summer rainfall over eastern China has significantly weakened since the late 1970s (Gao and Wang 2007; Xu et al. 2010; Zhu et al. 2007). Ye and Lu (2011) have explored the potential causes for the weakening of this relationship at a subseasonal scale and reported that ENSO-related rainfall anomalies are similar between early and late summer before the late 1970s; however, the anomalous rainfall patterns have almost reversed between early and late summer after the late 1970s. Ding et al. (2010) investigated the change of the relationship between the EASM and the tropical Indian Ocean (IO) from 1953–75 to 1978–2000 and attributed the EASM–IO relationship shift to the interdecadal change of the background state of the ocean–atmosphere system and the interaction between ENSO and the IO. Gao et al. (2014) uncovered a remarkable decadal shift in the relationship between spring AO and EASM on an interannual scale in the late 1990s and further indicated that a subtropical wave train from the North Atlantic Ocean to IO plays an important role in connecting AO and EASM in the post-1997 epoch, while the signal in the pre-1997 epoch is memorized and persists over the Pacific Ocean.

From the view of seasonal prediction, the predictor–predictand relationship in the IAM might change, as Goswami (2005) indicated it reflects the influence of the decadal variation on the interannual variation. The decadal change of predictor–predictand relationship influences the predictability of the statistical model; therefore, the statistical model must be constantly scrutinized and changed as necessary (Rajeevan et al. 2007). Here, two critical issues are highlighted when IAM is being constructed; they are (i) a revisit of the identification of precursors based on the latest data and (ii) a look at the development of the model in terms of the model training period. In this study, we aim to explore a new statistical prediction model in an effort to improve the seasonal prediction of YRV summer rainfall. The new model can 1) predict interannual and decadal rainfall, respectively, and 2) incorporate the changing relationship between precursor and interannual rainfall instead of the relationship being fixed as in previous models.

The framework of this study is structured as follows: Section 2 presents the data used in this work. Section 3 describes various methods for model development. Model development and prediction verification are presented in section 4. Section 5 provides a summary and brief discussion.

2. Data

Observed rainfall data for the period of 1961–2013 were obtained from China's 160-station monthly

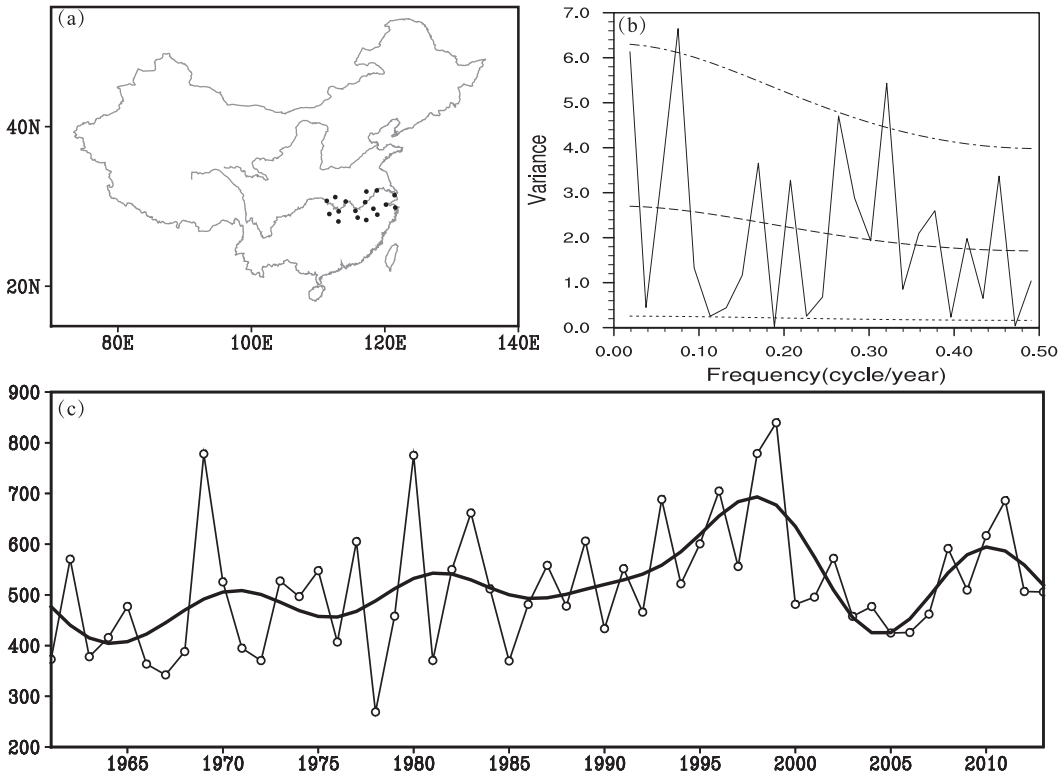

FIG. 1. (a) The 17 stations with rain gauges within the middle-lower reaches of the YRV (28°–32°N, east of 110°E). (b) Power spectrum for the YRV summer rainfall. Peaks above the upper dashed line indicate a confidence level > 90% against red noise. (c) Observed YRV summer rainfall (mm) from 1961 to 2013 and its decadal variation with a period > 9 yr (thick line).

rainfall dataset provided by the China Meteorological Administration. The total rainfall during June–August (JJA) averaged at 17 uniformly spread stations (Fig. 1a; Anqing, Changde, Changsha, Guixi, Hankou, Hangzhou, Hefei, Jiujiang, Nanchang, Nanjing, Ningbo, Shanghai, Quxian, Tunxi, Yichang, Yueyang, and Zhongxiang) within 28°–32°N and east of 110°E was calculated as YRV summer rainfall to be predicted.

For identifying precursors, we used monthly datasets of various parameters such as mean sea level pressure (SLP), 500-hPa geopotential height (H500), and sea surface temperature (SST). Monthly atmospheric data from March, April, and May were used, which were obtained from the Japan Meteorological Agency Japanese 55-year Reanalysis (JRA-55) dataset at 1.25° × 1.25° grid (Kobayashi et al. 2015). SST data were seasonal means of the preceding autumn (September–November), preceding winter (December–February), and spring (March–May), which were obtained from the monthly NOAA Extended Reconstructed SST, version 3b (ERSST.v3b), dataset at 2° × 2° grid (Smith et al. 2008).

To verify the capability of our statistical prediction model, 9-month-run ensemble forecasts by the NCEP Climate Forecast System, version 2 (CFSv2), covering a 13-yr period of 2001–13 from retrospective forecast (2001–10) and operational forecast (2011–13) were employed. CFSv2, the second version of the fully coupled dynamic seasonal forecast system, consists of the GFS at T126 resolution, the Modular Ocean Mode, version 4 (MOM4), at 0.25° × 0.5° grid coupled with a two-layer sea ice model, and the four-layer Noah land surface model. This system generated real-time seasonal forecasts since 30 March 2011 (Saha et al. 2014). The 9-month run is initiated every five days with four cycles of those days. For each month, the ensemble consists of 24 ensemble members with initial dates after the seventh of the previous month. As a matter of convenience, only the ensemble mean (equal weight mean of 24 ensemble members) was used here. Both 1-month-lead forecast (initiated in early May with initial dates from 11 April to 6 May) and 0-month-lead forecast (initiated in early June with initial dates from 11 May to 5 June) were employed in our study.

3. Methods

The spectrum analysis reveals two peaks in the YRV summer rainfall, with periods of 2–3 and 12–14 yr (Fig. 1b). The strong interannual and decadal variations,

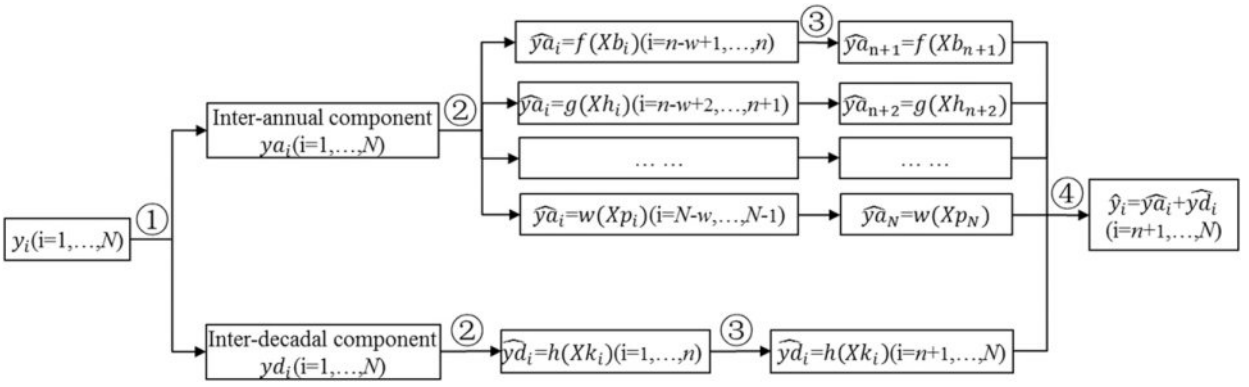

FIG. 2. Key stages in constructing the statistical prediction model: ① is time-scale decomposition; ② is construction of the IAM and DM; ③ is predictions made by IAM and DM; and ④ is combination of the predictions from IAM and DM to obtain the total rainfall.

shown in Fig. 1c, motivate us to predict the interannual and decadal rainfall, respectively, through developing distinct statistical prediction models.

Figure 2 shows the key stages of model development. The first step is time-scale decomposition; that is, the predictand, YRV summer rainfall, $y(t)$ is decomposed into the interannual component ya(t) (with a period $<$ 9 yr) and the decadal component yd(t) (with a period $>$ 9 yr) with Fourier decomposition filtering. We then construct both the IAM and DM. The entire data period of 1961–2013 ($N = 53$) is divided into a training period of 1961–2000 ($n = 40$) and an independent testing period of 2001–13 ($N - n$). For predicting decadal rainfall, a fixed model is trained with all 40-yr training data. For predicting interannual rainfall, the model is retrained year by year with the latest available data as prediction extends. Using all available data to train the IAM would be unlikely to produce the best predictions because the predictors in the IAM could change with the decadal variation of background state. What is the optimal length of the training period for a given IAM? We determined the optimal training period length w by comparing models trained with values of $w = 20$, 25, 30, 35, and 40 yr. Subsequently, statistical models are fitted with the latest w-yr data using the multilinear regression (MLR) method, and 13-yr independent predictions are conducted to validate the model's capability. A simple sum (with equal weight) of the predictions from the IAM and DM is the predicted total YRV summer rainfall.

In model construction, correlation analysis is first performed between rainfall and antecedent global atmosphere–ocean parameters. Any domain with a high (and significant) correlation coefficient is identified, and the area-weighted average value is calculated into an index as a potential predictor. Not every potential predictor is necessary in fitting the final regression equation. A "forward" stepwise regression screening nested with leave-one-out cross validation is utilized to select the optimal predictors from a set of potential predictors. For details of this cross-validation-based stepwise regression approach, see the appendix.

To measure the statistical model's prediction skill, statistics such as correlation coefficients, root-mean-square errors (RMSEs), and hit rates (the ratio of years in which the anomaly sign is predicted correctly to the total number of years) are employed. The bootstrap method (Stine 1985) is used to estimate the confidence intervals of our prediction.

4. Prediction of the YRV summer rainfall

The YRV summer rainfall is filtered into the interannual and decadal components, as are the nine antecedent global parameters (SLP and H500 in March, April, and May; SST in preceding autumn, preceding winter, and spring). The IAM and DM are then constructed, respectively.

a. Calibration and validation of the IAM

To determine the optimal training period length, the predicted interannual rainfall of 2001–07 with training period lengths of 20, 25, 30, 35, and 40 yr are examined (Fig. 3a). The model with a training period of 20 yr, for example, uses training data from 1981 to 2000 to predict rainfall of 2001, training data from 1982 to 2001 to predict rainfall of 2002, and so on. Figure 3b compares the prediction skill as measured with correlation coefficients and RMSEs. The model with a training period of 20 yr yields the best prediction, with the highest correlation coefficient of 0.63, the lowest RMSE of 25.8 mm (5.45% of the climatology), and a hit rate of 71.4%. Therefore, a 20-yr training period is chosen as the optimal length for IAM construction.

A 20-yr moving correlation between YRV summer rainfall and nine antecedent global parameters on

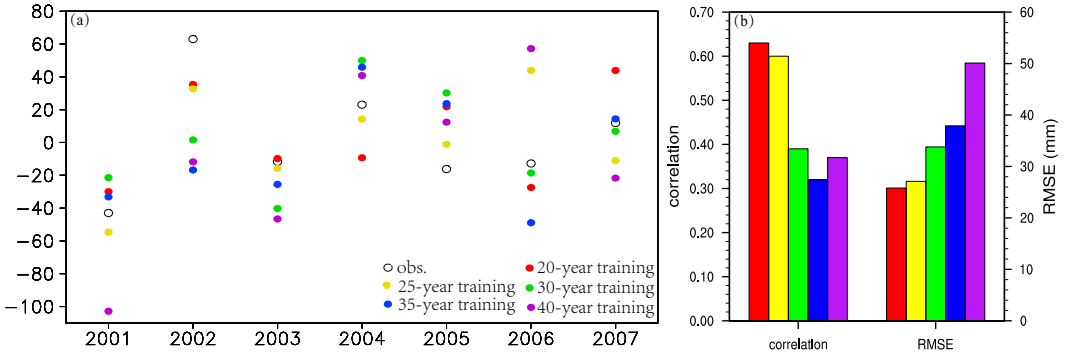

FIG. 3. (a) Interannual rainfall over 2001–07 from observations (open circles) and statistical prediction models (colored circles) trained with 20 (red), 25 (yellow), 30 (green), 35 (blue), and 40 (purple) yr of data. (b) Statistics for the 7-yr predictions.

an interannual scale are performed for the period of 1981–2012. As an example, the distribution of correlation coefficients between rainfall and May SLP is shown in Fig. 4. The locations of the significant correlation coefficients change over time: for example, a shrinking center with a positive correlation coefficient over eastern China and the northwestern Pacific and an expanding center with a negative correlation coefficient over western China and northern India. These time-changing features highlight the necessity for updating the prediction model using precursors reidentified with the latest data.

Because we were limited to illustrating all correlation distributions one by one (9 parameters for each model × 13 models for 2001–13), Fig. 5 shows four representative cases. It seems that some highly correlated signals, like ENSO, SST anomaly (SSTA) dipole over North Atlantic in Fig. 5a, and Mascarene high in Fig. 5b, have been indicated to influence the interannual variation of YRV summer rainfall by the previous studies (Pan 2005; Wu et al. 2009; Xue et al. 2003). Meanwhile, some highly correlated signals, like an east–west SLP anomaly (SLPA) dipole over East Asia in Fig. 5c and an east–west SSTA dipole over North Pacific in Fig. 5d, are newly identified. All these highly correlated signals are identified as potential precursors that were further selected through stepwise regression screening.

After stepwise regression screening, optimal precursors were selected to fit the final regression equation. Table 1 lists the prediction models fitted year by year for predicting rainfall of 2001–13, as well as the precursors used in each model. All prediction models, which consist of single and binary linear regression equations, are statistically significant at the 0.05 level. These models represent well the observed interannual rainfall in the fitting period as indicated by the lowest correlation coefficient of 0.73 (explaining 53% of interannual variance) for predicting rainfall of 2002 and the largest correlation coefficient of 0.92 (explaining 85% of interannual variance) for predicting rainfall of 2008. It is interesting that the precursors in the 13 models change every 4–5 years, generally from the preceding winter North Atlantic SSTA dipole for 2001 to the Mascarene high for 2002–05, followed by the East Asia SLPA dipole for 2006–10 and the preceding autumn North Pacific SSTA dipole for 2012/13.

Using the moving updated IAM, 13-yr independent predictions of the interannual YRV summer rainfall for 2001–13 were conducted year by year. Figure 6 shows the observed and statistically predicted interannual rainfall as well as the 95% confidence intervals. The IAM effectively captures the interannual variation of YRV summer rainfall with a correlation coefficient of 0.76 and an RMSE of 34.7 mm (7.32% of the climatology). It correctly predicts the rainfall anomaly signs of 10 out of 13 years for a hit rate of 76.9%. Moreover, the model can predict the amplitude of interannual variability well, as indicated by the interannual standard deviation ratio of 0.74 between prediction and observation.

b. Calibration and validation of the DM

The decadal correlations between YRV summer rainfall and nine antecedent global parameters were produced for the period of 1961–2000 (Fig. 7). The area with absolute of central correlation coefficient exceeding 0.6 is identified, and the area-weighted average is calculated as a potential precursor. All of the potential precursors are significantly correlated with the decadal variation of YRV summer rainfall at the 0.05 level after adjusting the degrees of freedom. After stepwise regression screening, only one precursor, the April SLPA in the central tropical Pacific (marked in Fig. 7e), was selected to fit the final regression equation given by

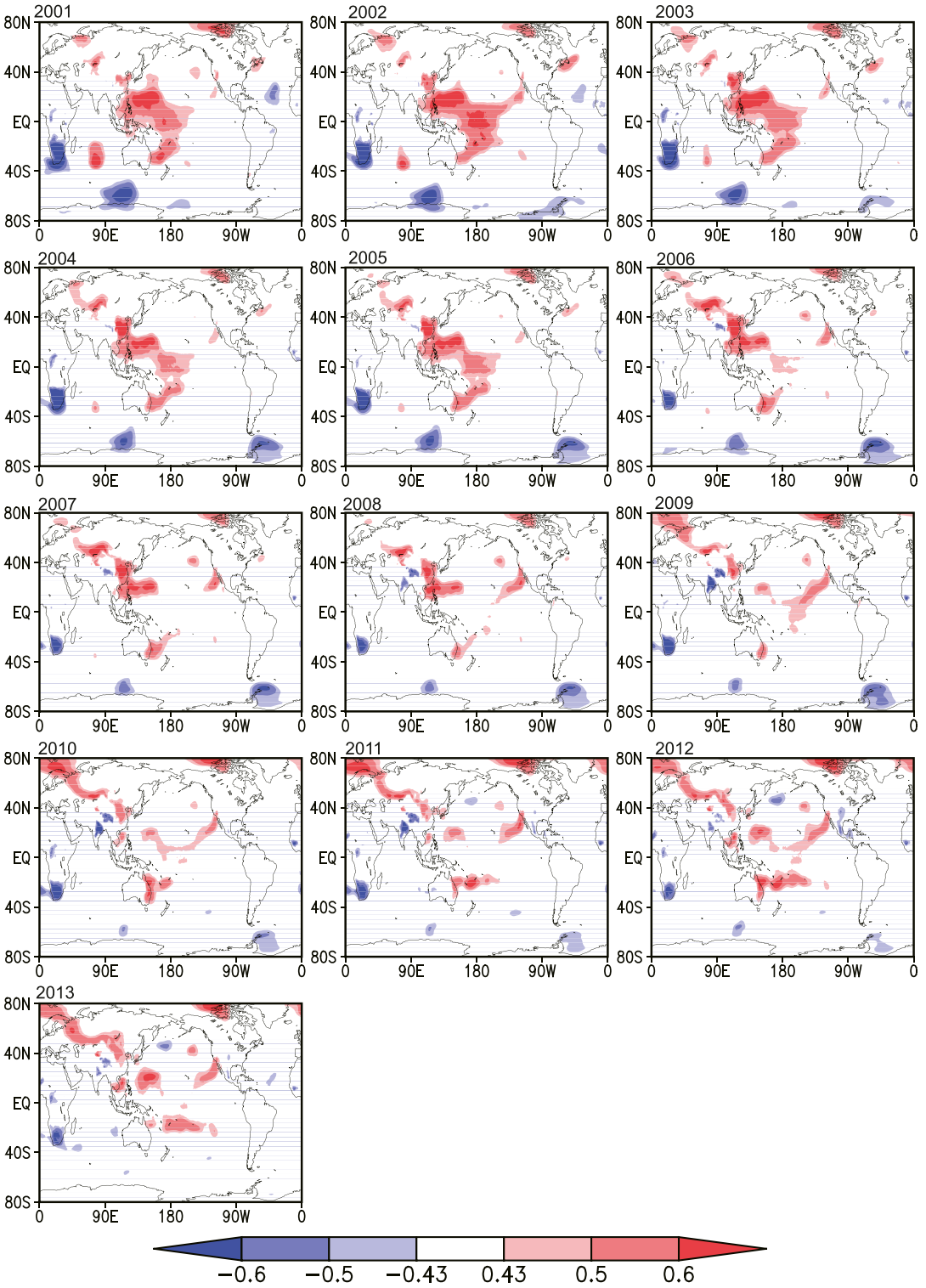

FIG. 4. The 20-yr moving correlations between YRV summer rainfall and SLP in May on an interannual scale over 1981–2012. The year on the top left of each panel is the year to be predicted. Color shading indicates statistical significance at the 0.05 level.

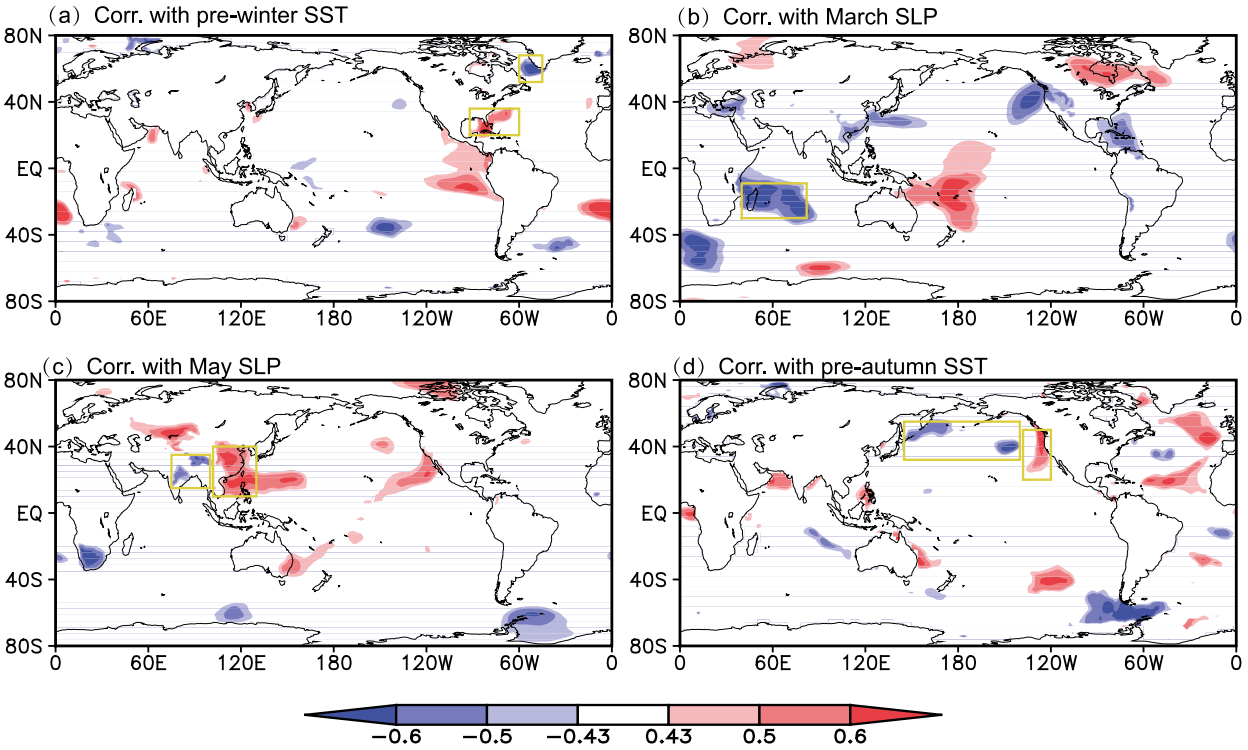

FIG. 5. Interannual correlations of YRV summer rainfall with (a) the preceding winter SST over 1981–2000, (b) March SLP over 1982–2001, (c) May SLP over 1988–2007, and (d) the preceding autumn SST over 1993–2012. Color shading indicates statistical significance at the 0.05 level.

$$\mathrm{yd}(t) = 516.8 + 55.9 \times \mathrm{SLP}_{\mathrm{CP}}(t), \qquad (1)$$

where $\mathrm{yd}(t)$ is the decadal rainfall at the tth year ($t = 1, \ldots, 40$) over 1961–2000, and $\mathrm{SLP}_{\mathrm{CP}}(t)$ is the tth value of the normalized $\mathrm{SLP}_{\mathrm{CP}}$.

The predictor $\mathrm{SLP}_{\mathrm{CP}}$ represents the central tropical Pacific SLPA in April. How does it affect the YRV summer (JJA) rainfall? Figures 8a–f show the decadal correlation of $\mathrm{SLP}_{\mathrm{CP}}$ with SST and 500-hPa vertical velocity in simultaneous April and ensuing May and summer (JJA). It seems that $\mathrm{SLP}_{\mathrm{CP}}$ is highly correlated with anomalous warming over both the western and the eastern tropical Pacific. The western anomalous warming enhances the Walker circulation in the western Pacific while the eastern anomalous warming reduces the Walker circulation in the eastern Pacific, as indicated by anomalous ascent over the western and eastern Pacific and anomalous decent over the central Pacific at the 500-hPa level. Anomalous zonal atmospheric circulation along the equator raises SLP over the central Pacific. It is noteworthy that this SSTA pattern with anomalous warming over the western and eastern tropical Pacific can persist from April to JJA, giving rise to influences on the summer climate. Figure 8g shows the decadal correlation of $\mathrm{SLP}_{\mathrm{CP}}$ with summer 850-hPa meridional winds. Corresponding to $\mathrm{SLP}_{\mathrm{CP}}$ positive anomalies, anomalous easterly and westerly flow prevails along the equator over the western and the eastern Pacific, and an anticyclonic anomaly appears over the western North Pacific as Rossby wave response. Abundant water vapor is transported to the YRV region along the western boundary of the enhanced western North Pacific anticyclone, resulting in sufficient YRV summer rainfall.

Figure 9 shows the observed and predicted decadal YRV summer rainfall from Eq. (1) in both the training period and the independent testing period as well as the 95% confidence intervals. The DM did a good job at capturing the decadal variation of rainfall in both the training period and the subsequent independent testing period. When compared with observation, the DM obtained a correlation coefficient of 0.76 and an RMSE of 48.4 mm (10.2% of the climatology) in the training period, and the high skill is maintained in the subsequent independent testing period, with a correlation coefficient of 0.77 and RMSE of 36.3 mm (7.7% of the climatology).

c. *Total rainfall*

It is straightforward to obtain the predicted total summer rainfall by summing the predictions from the IAM and the DM. Figure 10 shows the observed and

TABLE 1. The interannual rainfall models constructed year by year for predicting rainfall of 2001–13 and the precursors involved in each model (the "−1" indicates that the month(s) are associated with the preceding year).

Year	Model	Fitting scores		Precursors			
		Corr	RMSE (mm)	Name	Parameter	Month	Area
2001	$ya = -5.3 + 73 \times SST_{NAD}$	0.76	61.9	North Atlantic SST dipole (meridional)	SST	Dec (−1)–Feb	268°–300°E, 20°–36°N 300°–315°E, 52°–68°N
2002	$ya = 1.5 - 64.6 \times SLP_{MH}$	0.73	60.4	Mascarene high	SLP	Mar	40°–82°E, 30°–9°S
2003	$ya = 2.8 - 67.2 \times SLP_{MH}$	0.74	62.0	Mascarene high	SLP	Mar	40°–82°E, 30°–9°S
2004	$ya = -4.4 - 63.6 \times SLP_{MH}$	0.74	58.5	Mascarene high	SLP	Mar	40°–82°E, 30°–5°S
2005	$ya = -4.0 - 63.3 \times SLP_{MH}$	0.73	58.9	Mascarene high	SLP	Mar	40°–82°E, 30°–5°S
2006	$ya = 2.1 - 62.1 \times SLP_{EAD}$	0.76	53.5	East Asia SLP dipole (zonal)	SLP	May	85°–100°E, 25°–35°N 104°–130°E, 10°–40°N
2007	$ya = 2.8 - 52.3 \times SLP_{EAD} + 38.7 \times SST_{NAD}$	0.91	35.1	East Asia SLP dipole (zonal)	SLP	May	85°–100°E, 25°–35°N 104°–130°E, 10°–40°N
				North Atlantic SST dipole (meridional)	SST	Sep–Nov (−1)	308°–340°E, 46°–60°N 328°–350°E, 64°–72°N
2008	$ya = 0.5 - 55.4 \times SLP_{EAD} + 35.6 \times SST_{NAD}$	0.92	32.1	East Asia SLP dipole (zonal)	SLP	May	75°–100°E, 15°–35°N 102°–130°E, 10°–40°N
				North Atlantic SST dipole (meridional)	SST	Sep–Nov (−1)	308°–340°E, 46°–60°N 328°–350°E, 64°–72°N
2009	$ya = 2.8 - 45.6 \times SLP_{EAD} + 40.2 \times SST_{NPD}$	0.90	37.9	East Asia SLP dipole (zonal)	SLP	May	75°–100°E, 15°–35°N 102°–130°E, 10°–40°N
				North Pacific SST dipole (zonal)	SST	Mar–May	150°–180°E, 20°–34°N 200°–245°E, 5°–50°N
2010	$ya = -2.4 - 66.6 \times SLP_{EAD}$	0.82	46.3	East Asia SLP dipole (zonal)	SLP	May	75°–100°E, 15°–35°N 102°–130°E, 10°–45°N
2011	$ya = 3.5 + 45.5 \times SST_{NIO} + 46.4 \times SLP_{NE}$	0.86	40.4	SST in north Indian Ocean	SST	Sep–Nov (−1)	50°–90°E, 10°–25°N
				SLP in north Europe	SLP	May	0°–50°E, 60°–80°N
2012	$ya = 4.1 - 64.7 \times SST_{NPD}$	0.77	52.9	North Pacific SST dipole (zonal)	SST	Sep–Nov (−1)	145°–220°E, 32°–55°N 222°–240°E, 20°–50°N
2013	$ya = 5.1 - 63.3 \times SST_{NPD}$	0.78	50.9	North Pacific SST dipole (zonal)	SST	Sep–Nov (−1)	145°–220°E, 32°–55°N 222°–240°E, 20°–50°N

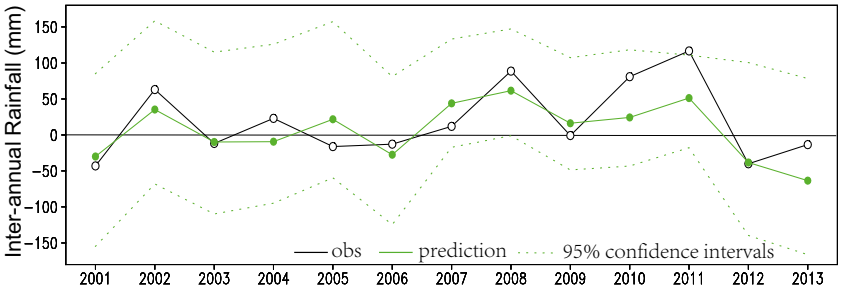

FIG. 6. Interannual rainfall over 2001–13 from observations (open circles) and statistical prediction model (green dots). Dashed lines denote the 95% confidence intervals.

statistical model predicted YRV summer rainfall anomalies for 2001–13 together with the 95% confidence intervals in comparison with the 1-month-lead and 0-month-lead CFSv2 forecasts. Because the prediction of summer rainfall is required in real time before June, we compared the statistical prediction with the 1-month-lead CFSv2 forecast. It seems that the 1-month-lead CFSv2 forecast can barely capture the observed rainfall variability with a correlation coefficient of 0.34 and an RMSE of 70.5 mm (14.9% of the climatology). By contrast, our statistical model provides skillful prediction, with correlation coefficient increased to 0.75 and RMSE reduced to 50.4 mm (10.7% of the climatology), even better than the best CFSv2 forecast initiated in early June (0-month-lead forecast), with a correlation coefficient of 0.62 and an RMSE of 58.3 mm (12.3% of the climatology). The 13-yr independent predictions demonstrate that our statistical prediction model has a high capability for capturing the YRV summer rainfall and is therefore promising for real-time seasonal prediction.

5. Summary and discussion

Because the YRV summer rainfall has such distinct variations on interannual and decadal scales and because

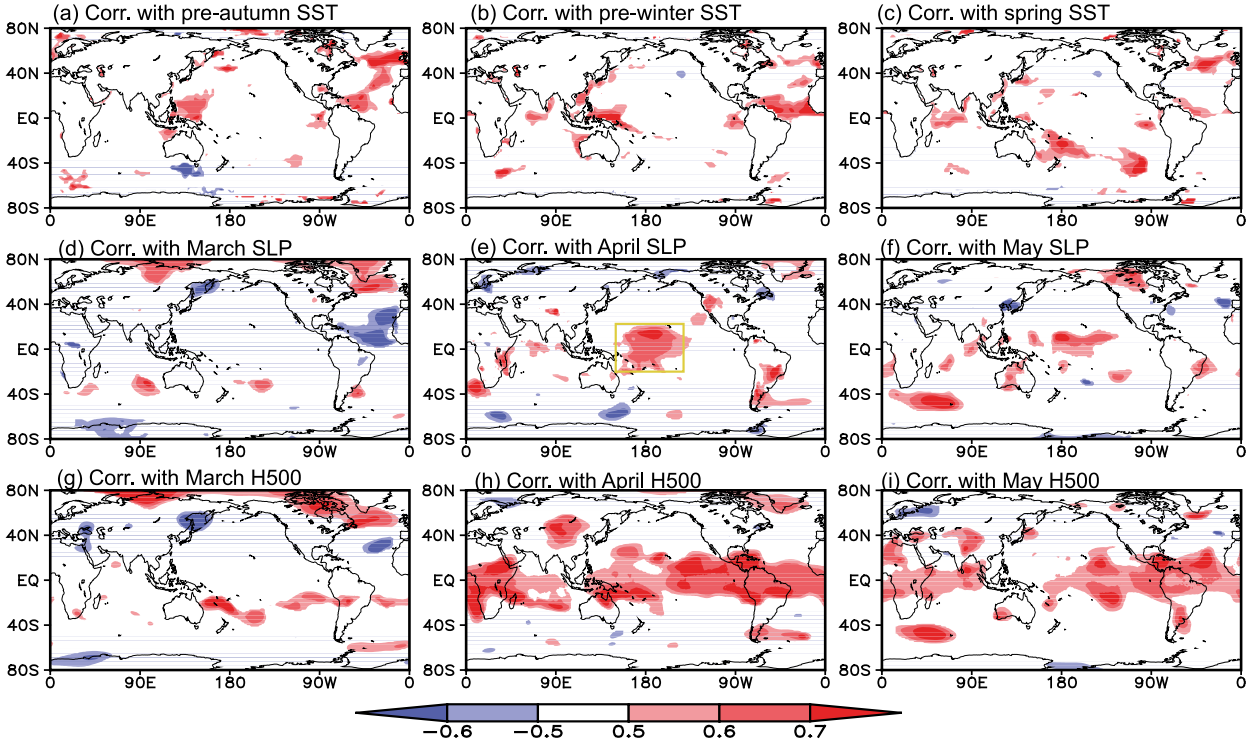

FIG. 7. Decadal correlations of YRV summer rainfall with (a)–(c) seasonal mean SST, (d)–(f) monthly SLP, and (g)–(i) monthly 500-hPa geopotential height over 1961–2000.

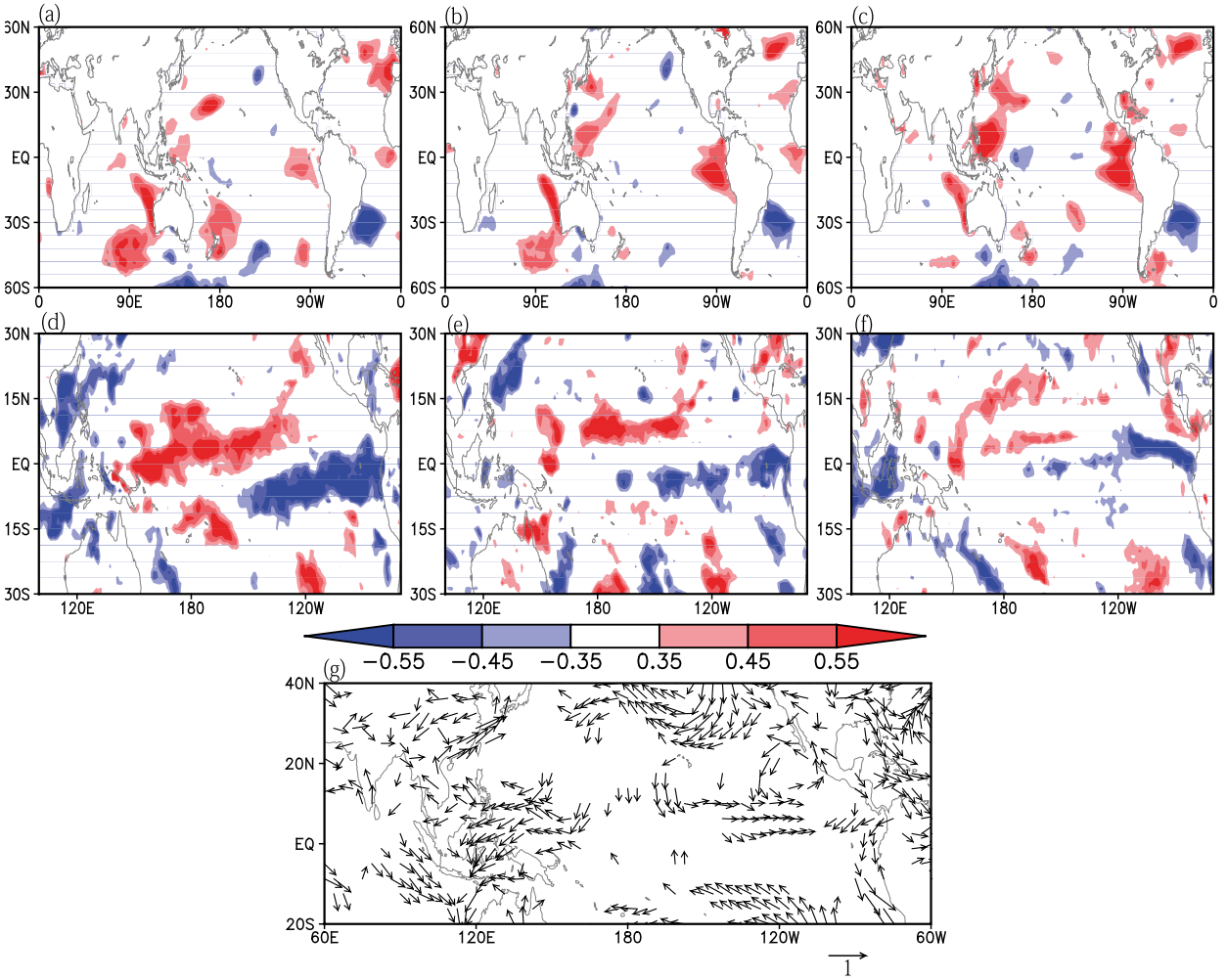

FIG. 8. Decadal correlations of SLP_{CP} with (a) April, (b) May, (c) JJA mean SST, (d)–(f) corresponding 500-hPa vertical velocity (Pa s^{-1}), and (g) JJA mean 850-hPa horizontal wind vectors (m s^{-1}) over 1961–2013.

the precursors that modulate the interannual rainfall change with the decadal variation of background state, the existing statistical prediction model, single and fixed, cannot provide stable and accurate prediction. Efforts continue toward this end, and attempts to adopt more suitable statistical techniques have resulted in the development of new models that make use of novel statistical ideas for improving the prediction skill for YRV summer rainfall.

In this study, we constructed the IAM and DM, respectively, with precursors selected based on a set of

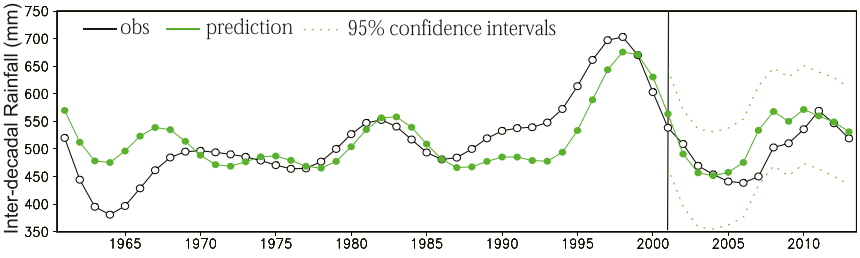

FIG. 9. Decadal rainfall over the training period of 1961–2000 and the independent testing period of 2001–13 from the observations (open circles) and statistical prediction model (green dots). Dashed lines denote the 95% confidence intervals.

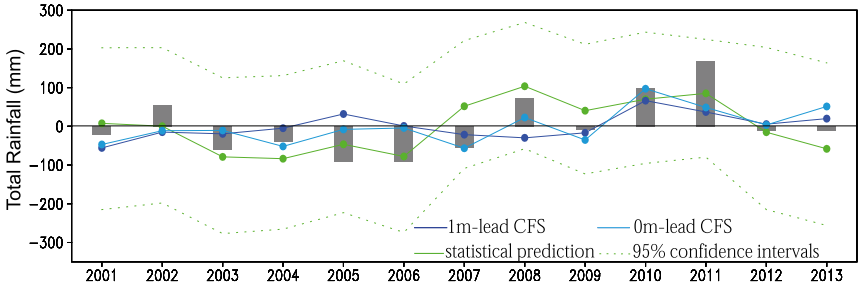

FIG. 10. YRV summer rainfall anomalies over 2001–13 from the observations (gray bars), 1-month-lead CFSv2 forecasts (dark blue), 0-month-lead CFSv2 forecasts (light blue), and the statistical prediction model (green). Dashed lines denote the 95% confidence intervals.

objective criteria. Moreover, the moving updating of the IAM with the latest data within an optimal length of training period partially offset the effect of decadal change of precursors that modulate the interannual rainfall. In IAM, an optimal training period length of 20 yr was determined. To predict the interannual rainfall of 2001–13 for validation, 13 regression models were fitted with precursors that change every 4–5 yr from the preceding winter North Atlantic SSTA dipole for 2001 to the Mascarene high for 2002–05, followed by the East Asia SLPA dipole for 2006–10 and the preceding autumn North Pacific SSTA dipole for 2012/13. This moving updated IAM demonstrated high skill in predicting interannual rainfall, with a correlation coefficient of 0.76 and a hit rate of 76.9%. The DM was linked to the April SLPA in the central tropical Pacific, and it maintained good performance in the testing period, with a correlation coefficient of 0.77 and an RMSE of 7.7%. Taking the predicted interannual and decadal rainfall together, we obtained the predictions of YRV summer rainfall for 2001–13. Our statistical model exhibited superior capability in predicting YRV summer rainfall even when compared with the best CFSv2 forecast initiated in early June (0-month-lead forecast), as indicated by the increased correlation coefficient from 0.62 to 0.75 and reduced RMSE from 12.3% to 10.7%.

Moving updating of the IAM cannot completely overcome the failure of statistical prediction stemming from the decadal change of precursors that modulate the interannual rainfall, especially for the rainfall of the year in which the decadal change occurs. It should be admitted that the moving updated model outperforms the traditional fixed statistical model because it can partially offset the effect of decadal change. When compared with those empirical models constructed previously to predict the EASM strength (Wu et al. 2009; Wu and Yu 2016), our statistical model generates comparable or even better skill, although differences exist in lead time and prediction period. In addition, comparisons with other dynamical predictions are needed since dynamical models are the main tools for future seasonal prediction. Li et al. (2016) evaluated an operational forecast system currently used in Met Office, Global Seasonal Forecast System, version 5 (GloSea5), whose ensemble mean has a general good prediction for YRV summer rainfall with a correlation coefficient of 0.55 for the 1992–2013 period. By comparison, our statistical model manifests superior skill for YRV summer rainfall that is promising for real-time operational prediction.

Acknowledgments. This study is jointly supported by the NSFC Project (41305053 and 41530424) and the National Programme on Global Change and Air–Sea Interaction (GASI-IPOVAI-03).

APPENDIX

Cross-Validation-Based Stepwise Regression Approach

The cross-validation-based stepwise regression (CVSR) approach is a "forward" stepwise screening procedure to select the "optimal" predictors from the potential predictor set. It employs leave-one-out cross validation in order to select the robust predictors and reduce the false possibility. The root-mean-square error between observation and cross-validation estimates (CV-RMSE) is taken as the criterion to evaluate the performance of potential predictors.

The CVSR method can be described in a general form using a series of iteration steps:

$$Y(t) = c + \sum_{i=1}^{p} \beta_i X_i(t) + \varepsilon(t), \qquad \text{(A1)}$$

where $Y(t)$ is the predictand for a $t = 1, \ldots, n$ year training period, $X_i(t)$ is the tth observation of the predictor X_i selected from candidate predictors $Z_1, \ldots, Z_m$

by the ith step in "forward" stepwise regression screening, c and β_i are model parameters, and $\varepsilon(t)$ is the error of the estimated model, Eq (A1). Model Eq. (A1) is established by $p < m$ steps:

Step 1: Regress the predictand $Y(t)$ onto each of the potential predictors $Z_{i_1}[i_1 \in (1, \ldots, m)]$ to obtain 1-predictor regression equation f_{i_1}. The performance of each 1-predictor regression equation is measured by CV-RMSE at step 1 as

$$\text{CV-RMSE}_{i_1} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\{Y(t) - f_{i_1,-t}[Z_{i_1}(t)]\}^2}, \tag{A2}$$

where regression equation $f_{i_1,-t}$ is fitted by $Z_{i_1}(j)[j \in (1, \ldots, n)\backslash(t)]$, that is, all observations excluding the tth one. If CV-RMSE_{i_1} is the smallest CV-RMSE achieved at step 1, that is, $\text{CV-RMSE}_1 = \min_{i_1 \in (1,\ldots,m)}(\text{CV-RMSE}_{i_1})$, then the potential predictor Z_{i_1} is selected as the first predictor, that is, $X_1 = Z_{i_1}$.

Step 2: Regress $Y(t)$ onto X_1 and each of the remaining $m - 1$ potential predictors $Z_{i_2}[i_2 \in (1, \ldots, m)]\backslash(i_1)]$, that is, all potential predictors except Z_{i_1}, to write 2-predictor regression equation f_{i_2}. The performance of each 2-predictor regression equation is measured by CV-RMSE at step 2 as

$$\text{CV-RMSE}_{i_2} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\{Y(t) - f_{i_2,-t}[X_1(t), Z_{i_2}(t)]\}^2}, \tag{A3}$$

where regression equation $f_{i_2,-t}$ is fitted by $X_1(j)$, $Z_{i_2}(j)[j \in (1, \ldots, n)\backslash(t)]$. Now, if CV-RMSE_{i_2} is the smallest CV-RMSE achieved at step 2, that is, $\text{CV-RMSE}_2 = \min_{i_2 \in (1,\ldots,m)\backslash(i_1)}(\text{CV-RMSE}_{i_2})$, and moreover, CV-RMSE_2 is significantly smaller than CV-RMSE_1, then the potential predictor Z_{i_2} is selected as the second predictor, that is, $X_2 = Z_{i_2}$; otherwise, stop selecting new predictors. To statistically test the significant reduction in CV-RMSE_2 relative to CV-RMSE_1, t and F tests are utilized to test the quadratic errors series between the observation and cross-validated estimates obtained at step 2 ⟨i.e., $\{Y(t) - f_{2,-t}[X_1(t), X_2(t)]\}^2, t \in (1, \ldots, n)$, where $f_{2,-t}$ is fitted by $X_1(j)$, $X_2(j)$, $[j \in (1, \ldots, n)\backslash(t)]$⟩ and at step 1 ⟨i.e., $\{Y(t) - f_{1,-t}[X_1(t)]\}^2$, $t \in (1, \ldots, n)$, where $f_{1,-t}$ is fitted by $X_1(j)$, $[j \in (1, \ldots, n)\backslash(t)]$⟩ in terms of the mean value and the variance.

Generally, at step k, assume that there are $k - 1$ predictors $X_1 = Z_{i_1}, \ldots, X_{k-1} = Z_{i_{k-1}}$ selected from the original potential predictors $Z_1, \ldots, Z_m$, and the associated smallest CV-RMSE at step $k - 1$ is

$$\text{CV-RMSE}_{k-1} = \min_{i_{k-1} \in (1,\ldots,m)\backslash(i_1,\ldots,i_{k-2})} \sqrt{\frac{1}{n}\sum_{t=1}^{n}\{Y(t) - f_{i_{k-1},-t}[X_1(t), \ldots, X_{(k-2)}(t), Z_{i_{k-1}}(t)]\}^2}, \tag{A4}$$

where regression equation $f_{i_{k-1},-t}$ is fitted by $X_1(j), \ldots, X_{(k-2)}(j), Z_{i_{k-1}}(j), [j \in (1, \ldots, n)\backslash(t)]$.

Regress $Y(t)$ onto $X_1, \ldots, X_{k-1}$ and each $Z_{i_k}[i_k \in (1, \ldots, m)\backslash(i_1, i_2, \ldots, i_{k-1})]$ of remaining $m - (k - 1)$ potential predictors to write k-predictor regression equation f_{i_k}. The performance of each k-predictor regression equation is measured by CV-RMSE at step k:

$$\text{CV-RMSE}_{i_k} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\{Y(\text{t}) - f_{i_k,-t}[X_1(t), \ldots, X_{(k-1)}(t), Z_{i_k}(t)]\}^2}, \tag{A5}$$

where regression equation $f_{i_k,-t}$ is fitted by $X_1(j), \ldots, X_{(k-1)}(j), Z_{i_k}(j), [j \in (1, \ldots, n)\backslash(t)]$.

If CV-RMSE_{i_k} is the smallest CV-RMSE achieved at step k, that is, $\text{CV-RMSE}_k = \min_{i_k \in (1,\ldots,m)\backslash(i_1,\ldots,i_{k-1})}(\text{CV-RMSE}_{i_k})$, and moreover, CV-RMSE_k is significantly smaller than CV-RMSE_{k-1}, then the potential predictor Z_{i_k} is selected as the kth predictor, that is, $X_k = Z_{i_k}$; otherwise, stop selecting new predictors. The t and F tests are utilized to statistically test the quadratic errors series between the observation and cross-validated estimates obtained at step k ⟨i.e., $\{Y(t) - f_{k,-t}[X_1(t), \ldots, X_k(t)]\}^2, t \in (1, \ldots, n) -$, where $f_{k,-t}$ is fitted by $X_1(j), \ldots, X_k(j), [j \in (1, \ldots, n)\backslash(t)]$⟩ and at step $k - 1$ ⟨i.e., $\{Y(t) - f_{k,-1,-t}[X_1(t), \ldots, X_{k-1}(t)]\}^2$, $t \in (1, \ldots, n)$, where $f_{k,-1,-t}$ is fitted by $X_1(j), \ldots, X_{k-1}(j)$, $[j \in (1, \ldots, n)\backslash(t)]$⟩ in terms of the mean value and the variance.

Finally, for all of the selected predictors via the CVSR procedure, the F test is used to test their regression coefficients. The insignificant predictors would be excluded, and the remaining predictors are used to fit the multilinear regression equation with the least squares method.

REFERENCES

Ding, R. Q., K. J. Ha, and J. P. Li, 2010: Interdecadal shift in the relationship between the East Asian summer monsoon and the tropical Indian Ocean. *Climate Dyn.*, **34**, 1059–1071, doi:10.1007/s00382-009-0555-2.

Fan, K., 2006: Atmospheric circulation in Southern Hemisphere and summer rainfall over Yangtze River valley (in Chinese). *Chin. J. Geophys.*, **49**, 599–606, doi:10.1002/cjg2.873.

——, H. J. Wang, and Y. J. Choi, 2008: A physically-based statistical forecast model for the middle-lower reaches of the Yangtze River valley summer rainfall. *Chin. Sci. Bull.*, **53**, 602–609, doi:10.1007/s11434-008-0083-1.

Gao, H., and Y. G. Wang, 2007: On the weakening relationship between summer precipitation in China and ENSO (in Chinese). *Acta Meteor. Sin.*, **65**, 131–137.

Gao, M. N., J. Yang, D. Y. Gong, and S. J. Kim, 2014: Unstable relationship between spring Arctic Oscillation and East Asian summer monsoon. *Int. J. Climatol.*, **34**, 2522–2528, doi:10.1002/joc.3849.

Gong, D. Y., J. Yang, S. J. Kim, Y. Gao, D. Guo, T. Zhou, and M. Hu, 2011: Spring Arctic Oscillation-East Asian summer monsoon connection through circulation changes over the western North Pacific. *Climate Dyn.*, **37**, 2199–2216, doi:10.1007/s00382-011-1041-1.

Goswami, B. N., 2005: The Asian monsoon: Interdecadal variability. *The Asian Monsoon*, B. Wang, Ed., Springer, 295–327.

Guo, Y., J. Li, and Y. Li, 2012: A time-scale decomposition approach to statistically downscale summer rainfall over north China. *J. Climate*, **25**, 572–591, doi:10.1175/JCLI-D-11-00014.1.

Kobayashi, S., and Coauthors, 2015: The JRA-55 Reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, doi:10.2151/jmsj.2015-001.

Lee, S. S., J. Y. Lee, K. J. Ha, B. Wang, and J. K. E. Schemm, 2011: Deficiencies and possibilities for long-lead coupled climate prediction of the western North Pacific–East Asian summer monsoon. *Climate Dyn.*, **36**, 1173–1188, doi:10.1007/s00382-010-0832-0.

Li, C. F., and Coauthors, 2016: Skillful seasonal prediction of Yangtze River valley summer rainfall. *Environ. Res. Lett.*, **11**, 094002, doi:10.1088/1748-9326/11/9/094002.

Liu, G., Q. Y. Zhang, and S. Q. Sun, 2008: The relationship between circulation and SST anomaly east of Australia and the summer rainfall in the middle and lower reaches of the Yangtze River (in Chinese). *Chin. J. Atmos. Sci.*, **32**, 231–241.

Nan, S. L., and J. P. Li, 2003: The relationship between the summer precipitation in the Yangtze River valley and the boreal spring Southern Hemisphere annular mode. *Geophys. Res. Lett.*, **30**, 2266, doi:10.1029/2003GL018381.

Pan, L. L., 2005: Observed positive feedback between the NAO and the North Atlantic SSTA tripole. *Geophys. Res. Lett.*, **32**, L06707, doi:10.1029/2005GL022427.

Ping, F., Z. X. Luo, and J. H. Ju, 2006: Differences between dynamics factors for interannual and decadal variations of rainfall over the Yangtze River valley during flood seasons. *Chin. Sci. Bull.*, **51**, 994–999, doi:10.1007/s11434-006-0994-7.

Rajeevan, M., D. S. Pai, R. A. Kumar, and B. Lal, 2007: New statistical models for long-range forecasting of southwest monsoon rainfall over India. *Climate Dyn.*, **28**, 813–828, doi:10.1007/s00382-006-0197-6.

Ruan, C. Q., and J. P. Li, 2016: An improvement in a time-scale decomposition statistical downscaling prediction model for summer rainfall over North China (in Chinese). *Chin. J. Atmos. Sci.*, **40**, 215–226.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate*, **21**, 2283–2296, doi:10.1175/2007JCLI2100.1.

Stine, R. A., 1985: Bootstrap prediction intervals for regression. *J. Amer. Stat. Assoc.*, **80**, 1026–1031, doi:10.1080/01621459.1985.10478220.

Wang, B., R. Wu, and X. Fu, 2000: Pacific–East Asian teleconnection: How does ENSO affect East Asian climate? *J. Climate*, **13**, 1517–1536, doi:10.1175/1520-0442(2000)013<1517:PEATHD>2.0.CO;2.

——, J. Liu, J. Yang, T. Zhou, and Z. Wu, 2009: Distinct principal modes of early and late summer rainfall anomalies in East Asia. *J. Climate*, **22**, 3864–3875, doi:10.1175/2009JCLI2850.1.

Wei, F. Y., 2006: Relationship between precipitation anomaly over the middle and lower reaches of the Changjiang River in summer and several forcing factors (in Chinese). *Chin. J. Atmos. Sci.*, **30**, 202–211.

Wei, Z. G., S. W. Luo, W. J. Dong, and P. J. Li, 1998: Snow cover data on Qinhai–Xizang Plateau and its correlation with summer rainfall in China (in Chinese). *Quart. J. Appl. Meteor.*, **9**, 39–46.

Wu, B. Y., L. G. Bian, and R. H. Zhang, 2004: Effects of the winter AO and the Arctic sea ice variations on climate variation over East Asia (in Chinese). *Chin. J. Polar Res.*, **16**, 211–220.

Wu, Z. W., and L. L. Yu, 2016: Seasonal prediction of the East Asian summer monsoon with a partial-least square model. *Climate Dyn.*, **46**, 3067–3078, doi:10.1007/s00382-015-2753-4.

——, B. Wang, J. P. Li, and F. F. Jin, 2009: An empirical seasonal prediction model of the East Asian summer monsoon using ENSO and NAO. *J. Geophys. Res.*, **114**, D18120, doi:10.1029/2009JD011733.

Xie, S. P., Y. Kosaka, Y. Du, K. M. Hu, J. S. Chowdary, and G. Huang, 2016: Indo-western Pacific Ocean capacitor and coherent climate anomalies in post-ENSO summer: A review. *Adv. Atmos. Sci.*, **33**, 411–432, doi:10.1007/s00376-015-5192-6.

Xu, T. T., P. W. Guo, J. Xie, and X. J. Yan, 2010: Interdecadal change of relationship between ENSO and the summer rain pattern of eastern China (in Chinese). *J. Nanjing Univ. Inf. Sci. Technol. Nat. Sci. Ed.*, **2**, 367–372.

Xue, F., H. J. Wang, and J. H. He, 2003: Interannual variability of Mascarene high and Australian high and their influences on summer rainfall over East Asia. *Chin. Sci. Bull.*, **48**, 492–497, doi:10.1007/BF03183258.

Ye, H., and R. Y. Lu, 2011: Subseasonal variation in ENSO-related East Asian rainfall anomalies during summer and its role in weakening the relationship between the ENSO and summer rainfall in eastern China since the late 1970s. *J. Climate*, **24**, 2271–2284, doi:10.1175/2010JCLI3747.1.

Zhu, Y. M., X. Q. Yang, X. Y. Chen, S. S. Zhao, and X. G. Su, 2007: Interdecadal variation of the relationship between ENSO and summer interannual climate variability in China (in Chinese). *J. Trop. Meteor.*, **23**, 105–116.

Zhu, Y. X., Y. H. Ding, and H. W. Liu, 2009: Simulation of the influence of winter snow depth over the Tibetan Plateau on summer rainfall in China (in Chinese). *Chin. J. Atmos. Sci.*, **33**, 903–915.