Atmosphere-Ocean

ISSN: 0705-5900 (Print) 1480-9214 (Online) Journal homepage: http://www.tandfonline.com/loi/tato20

Seasonal Prediction of the Yangtze River Runoff Using a Partial Least Squares Regression Model

Xiaochen Ye, Zhiwei Wu, Zhaomin Wang, Huying Shen & Jianming Xu

To cite this article: Xiaochen Ye, Zhiwei Wu, Zhaomin Wang, Huying Shen & Jianming Xu (2018) Seasonal Prediction of the Yangtze River Runoff Using a Partial Least Squares Regression Model, Atmosphere-Ocean, 56:2, 117-128, DOI: 10.1080/07055900.2018.1448751

To link to this article: https://doi.org/10.1080/07055900.2018.1448751

Published online: 23 Mar 2018.

Seasonal Prediction of the Yangtze River Runoff Using a Partial Least Squares Regression Model

Xiaochen Ye[1], Zhiwei Wu[2,5,*], Zhaomin Wang[3], Huying Shen[4], and Jianming Xu[5]

[1]*Earth System Modeling Center and Key Laboratory of Meteorological Disaster, Ministry of Education, Nanjing University of Information Sciences & Technology, Nanjing, Jiangsu 210044, People's Republic of China*
[2]*Institute of Atmospheric Sciences (IAS), Fudan University, Shanghai 200433, People's Republic of China*
[3]*College of Oceanography, Hohai University, Nanjing 210098, People's Republic of China*
[4]*Department of Hydrometeorological Forecasting, Bureau of Hydrology, Changjiang Water Resources Commission, Wuhan 430010, People's Republic of China*
[5]*Shanghai Key Laboratory of Meteorology and Health, Shanghai 200030, People's Republic of China*

[Original manuscript received 25 July 2017; accepted 22 December 2017]

ABSTRACT *As the longest river in Asia and the third-longest river in the world, the Yangtze River drains a large land area of the Eurasian continent. Seasonal prediction of the Yangtze River runoff is of crucial importance yet is a challenging issue. In this study, observed monthly runoff data are used to develop a new Yangtze River runoff index (YRI) for the 1950–2016 period. The YRI is not only able to quantify the runoff state of the Yangtze River but is also able to evaluate the intensity of the East Asian summer monsoon (EASM). The YRI is highly correlated with summer precipitation in the Yangtze River basin. It can also capture the principal components of the EASM circulation system.*

To predict the Yangtze River summer runoff, we employed a partial least squares (PLS) regression method to seek sea surface temperature (SST) modes in the previous winter associated with the YRI time series. The findings indicate that the first SST mode exhibits a strong link with the decaying phase of El Niño (or La Niña), while the second SST mode is related to a persistant mega-La Niña (or mega-El Niño). These suggest that an El Niño–Southern Oscillation (ENSO) or mega-ENSO may be an essential source of predictability for the Yangtze River summer runoff. After a 47-year training period (1950–1996), a physical-empirical PLS model is built then a 3-month-lead forecast is used to validate the model from 1997 to 2016. The PLS model exhibits promising predictive skill that is better than some state-of-the-art reanalysis data systems.

RÉSUMÉ *[Traduit par la rédaction] En tant que premier fleuve d'Asie et troisième fleuve du monde quant à sa longueur, le Yang Tsé draine une grande partie du continent eurasien. La prévision saisonnière du ruissellement du Yang Tsé est essentielle, mais pose un défi de taille. Nous utilisons ici les observations mensuelles de ruissellement afin de définir un nouvel indice de ruissellement du Yang Tsé pour la période de 1950 à 2016. Cet indice quantifie non seulement l'état du ruissellement du Yang Tsé, mais il évalue l'intensité de la mousson estivale de l'Asie de l'est. Il est fortement corrélé avec les précipitations estivales dans le bassin du Yang Tsé. Il représente également les principales caractéristiques de la circulation de la mousson estivale de l'Asie de l'est.*

Pour prévoir le ruissellement estival du Yang Tsé, nous avons effectué une analyse de régression partielle par les moindres carrés nous permettant de déterminer les modes de la température de surface de la mer (SST) de l'hiver précédent associés aux séries temporelles de l'indice. Les résultats indiquent que le premier mode de la SST possède un lien fort avec la phase de dissipation d'El Niño (ou de La Niña), tandis que le deuxième mode de la SST est lié à un épisode majeur et persistant de La Niña (ou d'El Niño). Ces résultats laissent penser que le phénomène El Niño-oscillation australe (ENSO) ou qu'un ENSO majeur serait un prédicteur essentiel pour le ruissellement estival du fleuve Yang Tsé. Après une période d'entraînement de 47 ans (1950 à 1996), nous construisons le modèle de régression physique-empirique, puis nous utilisons une prévision de 3 mois pour valider le modèle de 1997 à 2016. Le modèle de régression partielle fait preuve d'une capacité de prévision prometteuse, qui est supérieure à celle de certains systèmes de données de réanalyse de pointe.

KEYWORDS Yangtze River runoff; partial least squares regression; seasonal prediction

*Corresponding author's email: zhiweiwu@fudan.edu.cn

1 Introduction

The Yangtze River, being the longest river in China and the third-longest river in the world, has suffered frequent drought and flood events during the past decade. Basin-wide long-lasting catastrophic floods, with high water levels and a large flood volume, appeared in this basin during the summers of 1954 and 1998 (Yang & Wei, 1999). In 2006, a once-in-a-century drought event occurred in the Yangtze River valley, with a significant decrease in the total runoff volume in summer (Xu, He, Liu, & Chen, 2008). Such flood and drought disasters have caused huge economic loss and damage to property. Hence, quantifying the discharge variations in the Yangtze River basin and making long-lead-time predictions of the runoff during the flood season are critical.

Many previous studies have been conducted to investigate the characteristics and variation of the summer runoff in the Yangtze River basin (Dai & Zhu, 2010; Guo, Guo, Hou, Xiong, & Hong, 2015; Ju et al., 2011), but few studies have attempted to predict the interannual variations of the runoff directly. Because the interannual variation of the runoff is a highly complicated hydrological process influenced by many factors, most previous studies tended to first use dynamical models to predict the summer precipitation over the Yangtze River basin. The runoff was then calculated according to the empirical relationship between precipitation and runoff. Under these circumstances, the prediction skill of runoff is basically determined by that of precipitation. It is well known that dynamical models have noticeable biases in simulating summer precipitation (e.g., Wu & Li, 2008, 2009), particularly in the East Asian summer monsoon (EASM) regions (e.g., the Yangtze River basin). Even the most advanced dynamical models are barely able to accurately represent the complex physical processes of monsoon rainfall (e.g., Lee, Lee, Wang, Ha, & Jhun, 2013; Wang et al., 2008; Wu & Yu, 2015; Wu, Li, Jiang, & He, 2011). Up to now, there has been no dynamical model for the practical prediction of Yangtze River runoff (Lai, Jiang, Liang, & Huang, 2013).

Physical-empirical (PE) statistical models are considered an alternate method for dynamical prediction systems (Wu, Wang, Li, & Jin, 2009) and are established based on physical considerations rather than searching for statistical predictors (Wang et al., 2015). For instance, Wu et al. (2009) and Wu et al. (2011) established two PE models for the EASM and East Asian winter monsoon (EAWM), respectively. Lee et al. (2013) suggested that the leading EAWM modes are closely related to the El Niño–Southern Oscillation (ENSO) and are far more predictable by statistical models than other modes. However, during the construction of the aforementioned linear PE model, the collinearity among the dominant factors would make the regression coefficients unreliable or lead to an inaccurate prediction. In light of these, PE models are usually constructed using a partial least squares (PLS) regression (e.g., Wu & Yu, 2015 among others). The PLS regression model can eliminate multicollinearity well with an ordinary least squares strategy.

In this study, we attempt to use the PLS regression model to predict the interannual variations of the Yangtze River runoff during flood season and try to answer the following questions. Is there any possibility of constructing an appropriate and reasonable index to quantify the interannual runoff variations? If so, what is the physical basis for this index? Is there an apparent connection between the runoff variations and sea surface temperature (SST) in the preceding season? How close are the hindcast results to reality and so can be used for real-time runoff predictions?

To answer these questions, a new runoff index is developed using real-time observational runoff data from the Yangtze River basin and its physical basis is clarified (Section 3). In Section 4, the previous winter (December, January, and February (DJF)) SST patterns are identified as the predictive factors with the PLS method. A PE PLS model is constructed for the prediction of interannual variations of the runoff in the Yangtze River flood season using DJF SST conditions. The relationship between the Yangtze River runoff index (YRI) and the EASM is discussed in Section 5. Finally, conclusions and a discussion are provided in Section 6.

2 Data, model, and methodology

a *Data*

The main datasets employed in this study include (i) monthly mean precipitation data from the global land precipitation (PREC/L) data for the 1950–2016 period gridded at 1.0° × 1.0° resolution (Chen, Xie, Janowiak, & Arkin, 2002); (ii) monthly mean atmospheric fields, taken from the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR) reanalysis, available at 2.5° × 2.5° resolution for the 1950–2016 period (Kalnay et al., 1996); (iii) monthly mean SST data from the improved Extended Reconstructed SST, version 3 (ERSST V3), gridded at 2.0° × 2.0° resolution for the 1949–2016 period (Smith, Reynolds, Peterson, & Lawrimore, 2008); (iv) monthly mean runoff reanalysis data, taken from the National Oceanic and Atmospheric Administration–Cooperative Institute for Research in Environmental Sciences (NOAA–CIRES) Twentieth Century Reanalysis, version 2c (20CR V2c) for the 1950–2014 period (Compo et al., 2011); (v) monthly mean runoff data at three major hydrological stations (i.e., Yichang, Hankou, and Datong) for the 1950–2016 period; and (vi) monthly precipitation from 42 representative meteorological stations distributed in the Yangtze River basin (Fig. 1). The data were downloaded from the website of the China Meteorological Administration (http://cmdp.ncc-cma.net/nccdownload/index.php?ChannelID=1) and cover the 1951–2016 period. In this study, winter refers to December, January, and February (December of the previous year, January and February in the current year), spring refers to March, April, and May (MAM), summer refers to June, July, and August (JJA), and autumn refers to September, October, and November (SON).

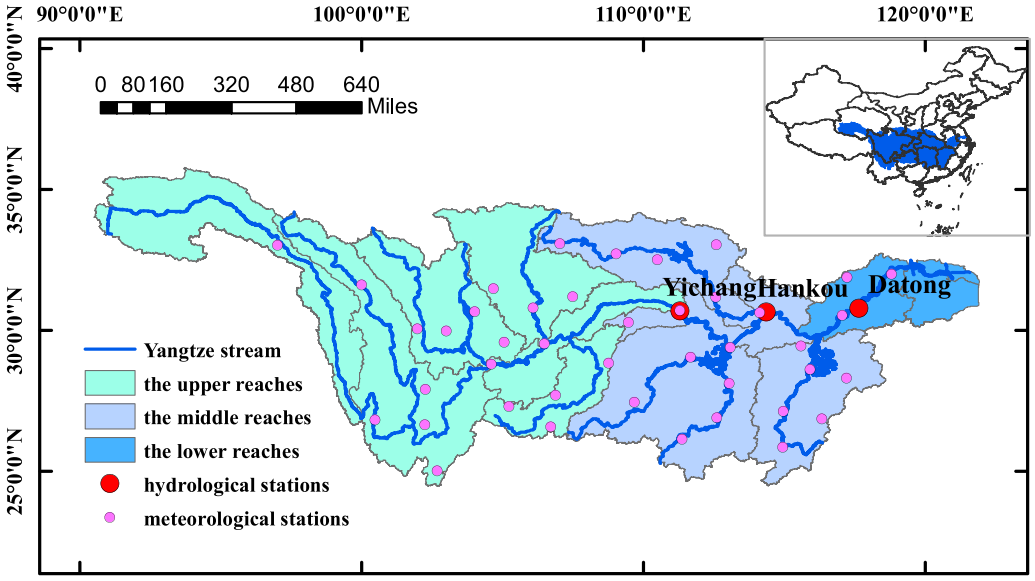

Fig. 1 Map of the Yangtze River basin showing the mainstream and distribution of key gauging stations. Red dots represent the hydrological stations along the Yangtze River and purple dots represent the meteorological stations in the Yangtze River basin.

b *Location of the Yangtze River Basin and the Observation Stations*

The Yangtze River is the longest river in China, which can be traditionally divided into three reaches (Fig. 1). Yichang station is located at the intersection point of the upper and middle reaches, representing the upstream runoff. Hankou station is situated midstream, with the Hanjiang River passing through the station. Datong station is the monitoring station for the lower Yangtze catchment area, taking up about 95% of the drainage area of the Yangtze basin, representing the total river discharge. Therefore, runoff at these three gauging stations can generally represent the main features of the entire catchment area (Zhao, Li, Dai, Zhang, & Yan, 2010), especially the river runoff measured at the estuary (Datong station is generally regarded as the representative station for the entire basin), which reflects changes in the entire Yangtze River basin.

c *Methodology*

For seasonal prediction, a PE PLS regression model was established. Assume that we intend to predict a time series Y_i (with subscript i representing the year, $i = 1$ to n). First, we should select a predictor field with a physical basis, which has m grids with the same time series of n years. Thus, the predictor field can be written as an $n \times m$ matrix ($\mathbf{X}_{ij}$ at each grid, $i = 1$ to n, j, representing the grids, $j = 1$ to m). Using regression analysis, the principal components a_k and b_k (k is the regression time, $k = 1$ to k) were extracted from $\mathbf{X}_{ij}$ and $\mathbf{Y}_i$, respectively. That is to say, a_k is a linear combination of $\mathbf{X}_{ij}$, and b_k is a linear combination of $\mathbf{Y}_i$. During the modelling process, a_k and b_k should be able to carry as much information as possible, and the correlation between a_k and b_k should reach a maximum. These two requirements indicate that a_k and b_k should represent the data $\mathbf{X}_{ij}$ and $\mathbf{Y}_i$ as well as possible, and the independent component a_k can maximize the variance explained in the dependent variable b_k. To obtain the time series of a_k and b_k, we should repeat this step on the residual matrices if necessary. Finally, if k principal components of $\mathbf{X}_{ij}$ are extracted, the equation between a_k and the predictand $\mathbf{Y}_i$ can be written as

$$\mathbf{Z}_{ik} = a_1\mathbf{X}_{i1} + a_2\mathbf{X}_{i2} + \cdots + a_k\mathbf{X}_{ij},$$
$$\mathbf{Y}_i = b_1\mathbf{Z}_{i1} + b_2\mathbf{Z}_{i2} + \cdots + b_k\mathbf{Z}_{ik}.$$

And the final equation between $\mathbf{X}_{ij}$ and the predictand $\mathbf{Y}_i$ is

$$\mathbf{Y}_i = w_1\mathbf{X}_{i1} + w_2\mathbf{X}_{i2} + \cdots + w_j\mathbf{X}_{ij}.$$

We refer the reader to Wu, Lin, Li, and Tang (2013) and Wu and Yu (2015) for details of the PLS.

This statistical method can explain both runoff and SST variance and has the advantage of being able to eliminate multi-collinearity. The regression model is allowed under the condition that the number of samples is less than the number of independent variables. In addition, the PLS regression contains all the original variables in the final model, which makes it simple to identify the system's information and noise, and it will be easier to explain the regression coefficients of the independent variables. Another advantage of the PLS method is that it is simple to calculate, has high precision, and is easy to interpret qualitatively.

We also compare the prediction results with 20CR V2c data to test the performance of our model.

In order to find the relationship between the YRI and water vapour transport, we use the following equations to compute qu and qv, the column-integrated horizontal water vapour

fluxes in the zonal and meridional directions, respectively (Zhao et al., 2016):

$$qu = \frac{1}{g}\int_{p_s}^{p_t} qudP,$$

$$qv = \frac{1}{g}\int_{p_s}^{p_t} qvdP,$$

where g is the acceleration due to gravity, u and v are the zonal and meridional wind components, respectively, q is specific humidity, P is pressure, p_s is surface pressure, and p_t is the pressure at the top of the atmosphere. To elucidate the relationship between the YRI and column-integrated water vapour transport, we then calculate the correlation coefficients.

3 Definition of the Yangtze River runoff index

The runoff formation process is a complex physical process through all aspects from precipitation on the surface of the basin to water flow convergence to the basin outlet section, which can be summarized as the process of runoff generation and confluence. Given that the occurrence of floods and droughts is dependent on the total runoff volume, we calculated the runoff data series at the three key stations (Yichang, Hankou, and Datong) along the Yangtze River from 1950 to 2016 to identify the concentration period of the runoff. The 67-year average monthly runoff distribution at Yichang, Hankou, and Datong station are shown separately for 12 months as bar charts in Fig. 2. Results show that runoff in the Yangtze River is mainly concentrated in the June to September period (JJAS) at all three stations. Table 1 lists the correlation coefficients among runoff variability over the three hydrological stations. The correlation coefficients are computed using monthly mean data during the boreal summer from June to September for the 67-year period from 1950 to 2016. There is a strong correlation among the runoffs at all three hydrological stations, which shows consistency in the variability of runoff at the upper, middle, and lower reaches of the Yangtze River.

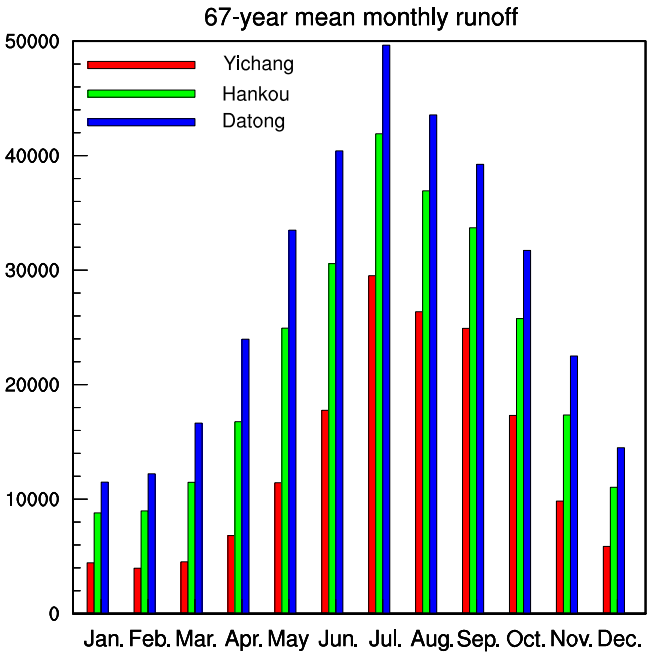

Fig. 2 The 67-year mean monthly runoff ($m^3\ s^{-1}$) at hydrological stations from 1950 to 2016. The red, green, and blue bars represent the Yichang, Hankou, and Datong stations, respectively.

Given the consideration of the runoff concentration period in the Yangtze River basin, we defined a new runoff index, YRI, using the normalized JJAS mean runoff data for Datong station (at the estuary of the Yangtze River) for the 1950–2016 period (Fig. 3). Monthly precipitation from 42 representative meteorological stations distributed in the Yangtze River basin (Fig. 1) were chosen to establish the regional mean precipitation series. The locations of the 42 meteorological stations are shown in Fig. 1. The mean summer rainfall at the 42 stations is shown in Fig. 3. The correlation coefficient between the YRI and rainfall averaged over the Yangtze River is 0.90 for 1951–2016.

4 Seasonal prediction based on the PLS model

a *Climate Element Field and Physical Precursor*

To uncover the physical basis for this index, we first review the climatological mean circulation features during the flooding season. Figure 4 depicts the correlation between the YRI and the seasonal mean rainfall in East Asia during summer. In the Yangtze River basin, there is an obvious positive correlation between the YRI and JJA rainfall. The YRI also shows a good relationship with precipitation in the Yellow River basin, suggesting that this index is a useful indicator for summer precipitation in southeastern China.

Meanwhile, Fig. 5 shows the correlation patterns of geopotential height and winds in summer with respect to the YRI for the 1950–2016 period at 200 hPa (Fig. 5a), 500 hPa (Fig. 5b), and 850 hPa (Fig. 5c). At 200 hPa, a large-scale anticyclonic circulation extends from the Indian Peninsula to southern China (Fig. 5a), which corresponds to the South Asian high (SAH) system. The subtropical westerly jet exists on the north flank of the SAH, which is a crucial system to the formation and maintenance of the Meiyu rainband (Sampe & Xie, 2010). The correlation pattern for 500 hPa shows

TABLE 1. Correlation coefficients among runoff variability over three hydrological stations (i.e., Yichang, Hankou, and Datong).

	Yichang	Hankou	Datong
Yichang	1.00	0.85**	0.69**
Hankou	0.85**	1.00	0.92**
Datong	0.69**	0.92**	1.00

Notes: Calculations are based on monthly means from June to September for the 67-year period from 1950 to 2016. Double asterisks (**) represent coefficients that are significant above the 99% confidence level.

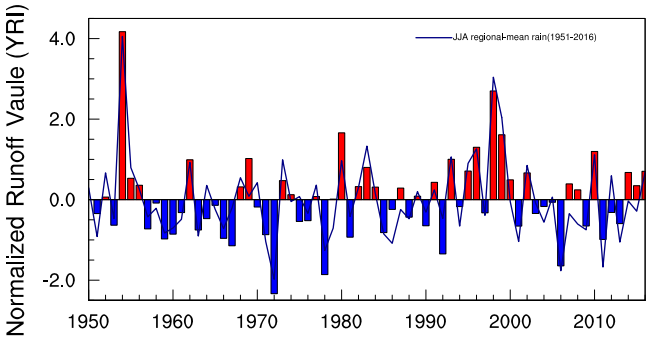

Fig. 3 Time series of normalized Yangtze River runoff index (YRI) and the regional mean rainfall.

significant positive anomalies over East Asia (Fig. 5b), which is linked to the subtropical high (SH). At 850 hPa, an anomalous cyclone is seen near the region of Bohai and the Yellow Sea, and an anomalous anticyclone exists around Southeast Asia (Fig. 5c). The convergence of anomalous northerly and southerly winds results in more moisture in the Yangtze River basin and favours more rainfall, thus generating more runoff.

Our study also indicates that there is a statistically significant relationship between the runoff in the Yangtze River basin and the strength of the column-integrated moisture fluxes (*qu* and *qv*) during summer. Along with the 850 hPa wind field in Fig. 5, analysis of correlations between the YRI and column-integrated moisture fluxes (Fig. 6) shows signs of moisture convergence (negative divergence) in the Yangtze River valley. When the flux convergence in the column-integrated moisture transport in the Yangtze River basin is enhanced, moisture transport originating from the low latitude ocean shows two opposing patterns and moisture transport originating from the high latitude ocean shows moisture divergence. This pattern can provide sustained abundant moisture to facilitate the development of extreme runoff in this basin. The enhanced moisture flux convergence is, therefore, a key factor in understanding extreme runoff in the Yangtze River basin.

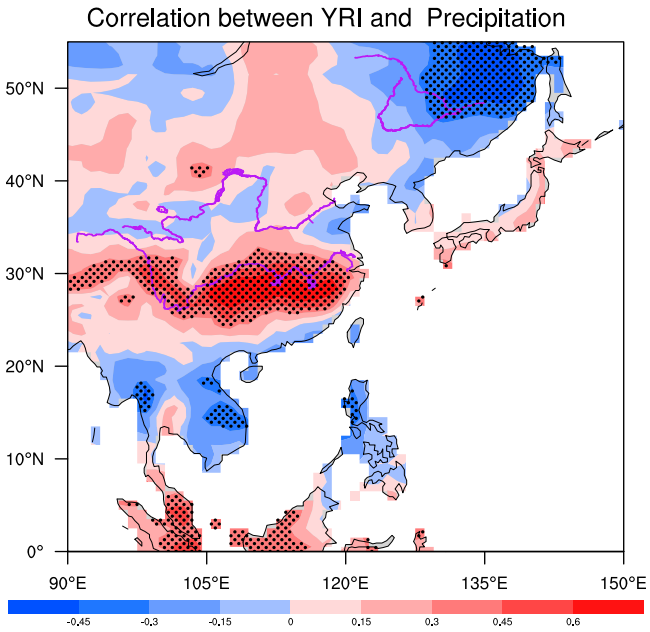

Fig. 4 Correlation between YRI and June, July, and August (JJA) mean precipitation. Black dotted areas represent correlation coefficients significant above the 99% confidence level.

By comparing the correlation pattern of the YRI with the SST from the preceding winter with the one from the autumn (based on a three-month mean) (Fig. 7), distinctly different correlation patterns can be seen in the equatorial central to eastern Pacific. A negative correlation develops (positive correlation decays) from the previous winter to the following autumn. While in summer, a significant negative correlation extends westward into the central Pacific. The SST pattern corresponds to a transition from an El Niño pattern during the previous winter to a La Niña pattern in the subsequent summer.

The above discussion indicates that the YRI may provide an effective indicator for characterizing the runoff variation in the Yangtze River basin, and the SST pattern preceding the flooding season may provide a physically meaningful predictor for this index.

b *Winter SST Patterns Preceding the Flooding Season*

To determine the causative physical connection between the predictor SST and the YRI, we use PLS regression to identify the SST patterns in the preceding winter. We regard the runoff index series as the dependent variable, and the independent variable is linked to an entire predictor field, which is the SST of the preceding winter over the oceans (0°–358°E, 60°S–60°N) because the correlation between the YRI and SST in DJF (Fig. 7a) is higher than in MAM (Fig. 7b). In building this regression model, the two values of the SST modes represent the importance of the SST conditions in capturing the variation of the YRI index and the total SST variance. The first two PLS modes emerge as statistically significant, which together explain 42.24% of the variance of the SST and 38.26% of the YRI. Figure 8 shows the first two SST modes of the preceding winter (Fig. 8a and 8c) and the corresponding time series (Fig. 8b and 8d) from 1950 to 1996. The first mode closely resembles the SST pattern associated with ENSO. Significant positive correlations of 2.0–4.0 can be seen over the equatorial central-eastern Pacific, while negative correlations are observed over the northern and southern Pacific. This suggests that warming in the equatorial Pacific may strengthen the runoff. The second mode appears to be a pattern related to a mega-La Niña, and its residual time series (Fig. 8d) exhibits a very weak downward trend. The definition of a mega-El Niño–Southern Oscillation (mega-ENSO) index includes both the tropical Pacific and its subtropical extensions. With a broader spatiotemporal scale that differs from the conventional ENSO, the mega-ENSO

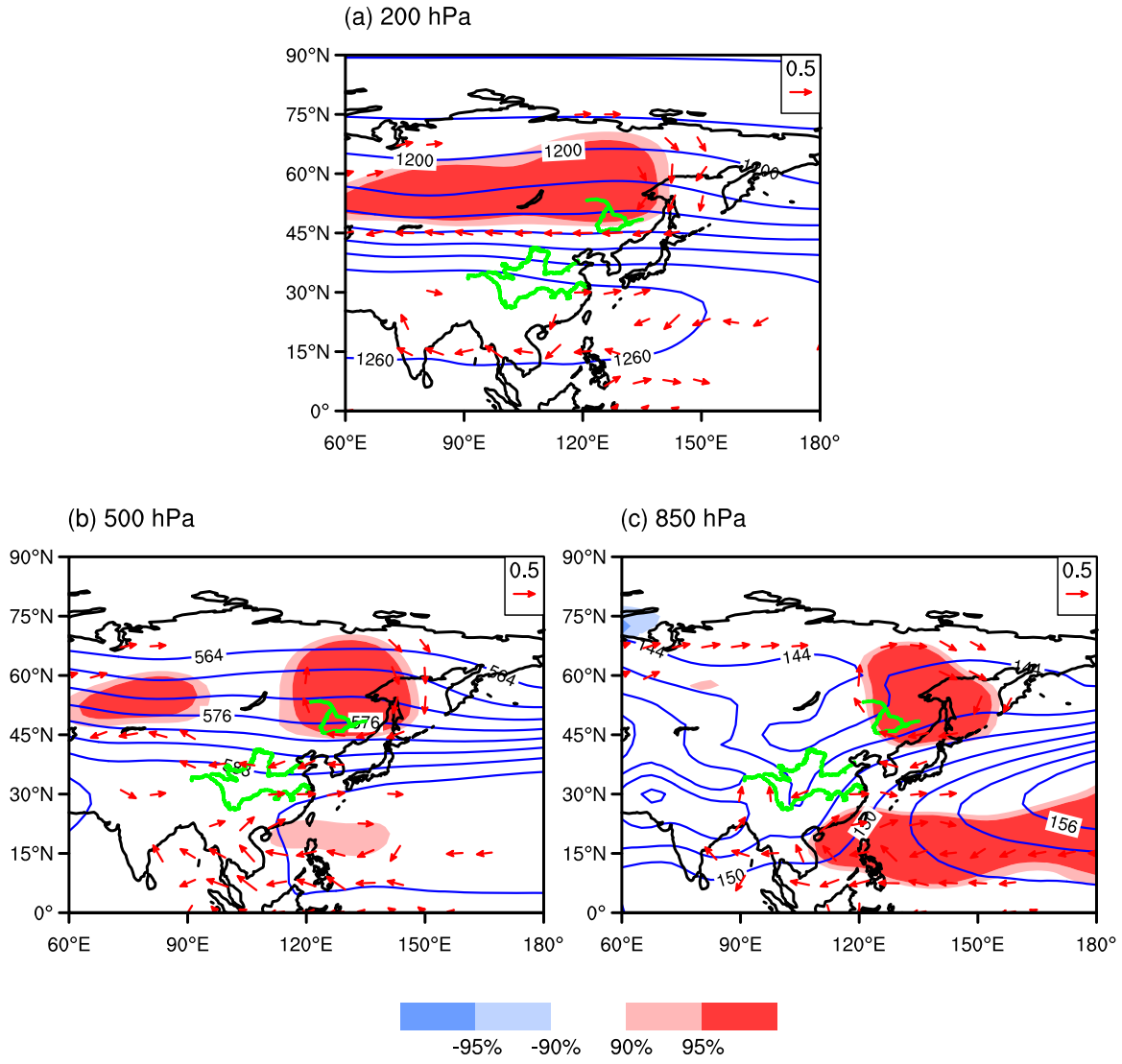

Fig. 5 Correlation maps at (a) 200 hPa, (b) 500 hPa, and (c) 850 hPa. Shading indicates the correlation coefficients between the YRI and JJA heights. Only those significant above the 95% and 90% confidence levels based on a Student's *t*-test are plotted. Vectors show the relationships between the YRI and the winds (m s^{-1}). Only winds significant at the 95% confidence level are plotted. Contours represent the summer climatological mean of the heights (gpm).

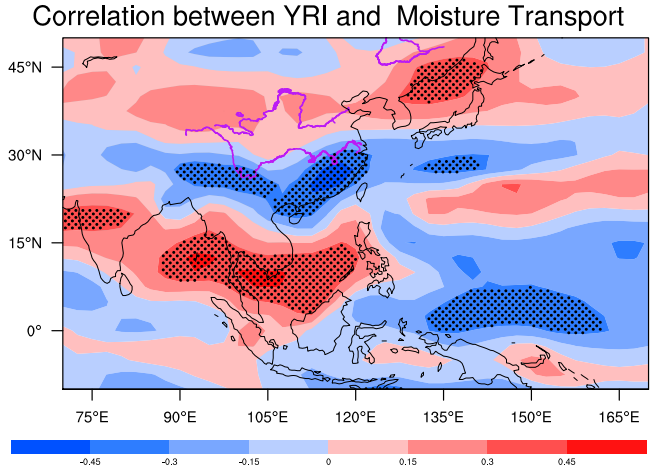

Fig. 6 Correlations between the YRI and column-integrated moisture fluxes.

index may contain more multi-time-scale information than a conventional ENSO index (Wang et al., 2013).

The first two PLS modes corresponding to the seasonal evolution of SST patterns from the winter to the following autumn are shown in Fig. 9. The first PLS mode clearly represents the decaying phase of ENSO, which is characterized by positive anomalies in the western Pacific and negative anomalies in the eastern tropical Pacific in the preceding winter. Considering the time scales and spatial distribution of the second mode, the second mode is related to a mega-La Niña pattern in the Pacific (Wang et al., 2013). This pattern remarkably persists for almost an entire year. The stability may reflect its superior predictive power.

Taking the moderate correlation coefficient between the YRI and SST conditions into account (Figs 7 to 9), we will further study the use of the preceding DJF mean SST field

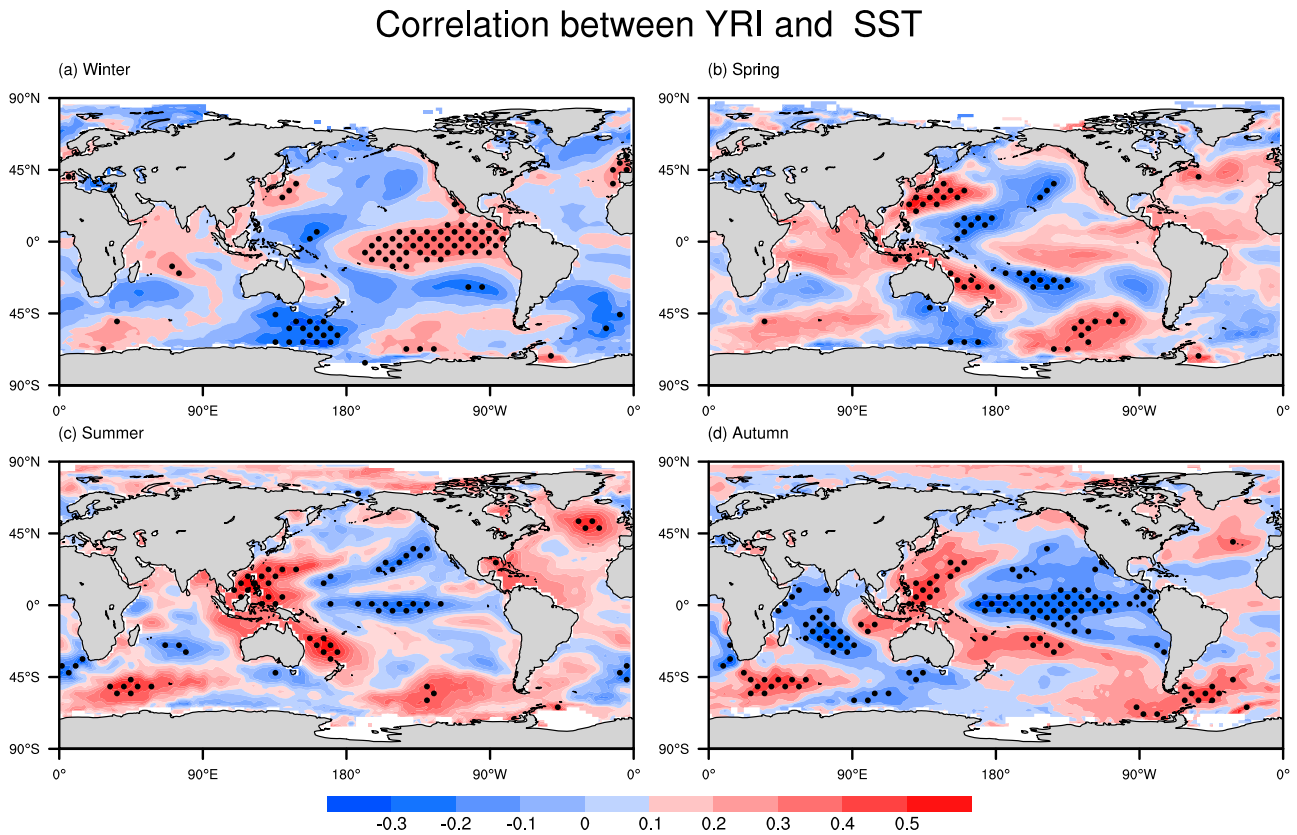

Fig. 7 Correlation maps between the YRI and SST from the preceding winter to the following autumn (based on three-month means). The black dots indicate correlation coefficients significant above the 95% confidence level.

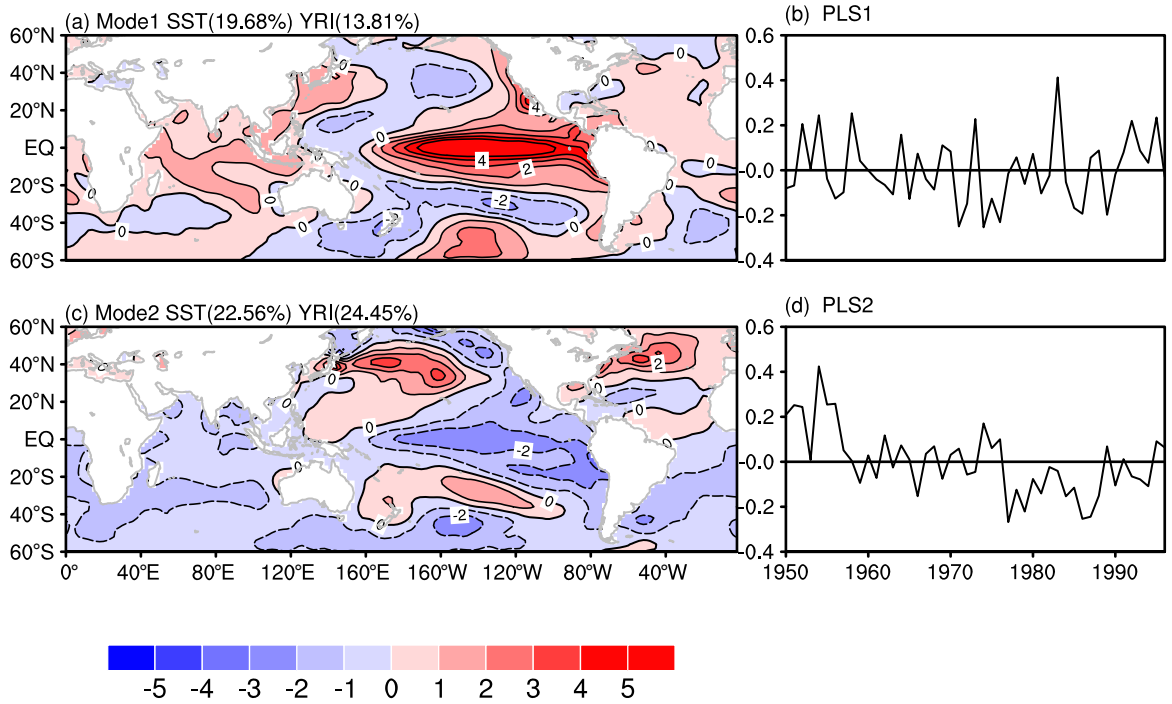

Fig. 8 The two leading SST modes and time series in the preceding winter from the results of the partial least squares (PLS) regression.

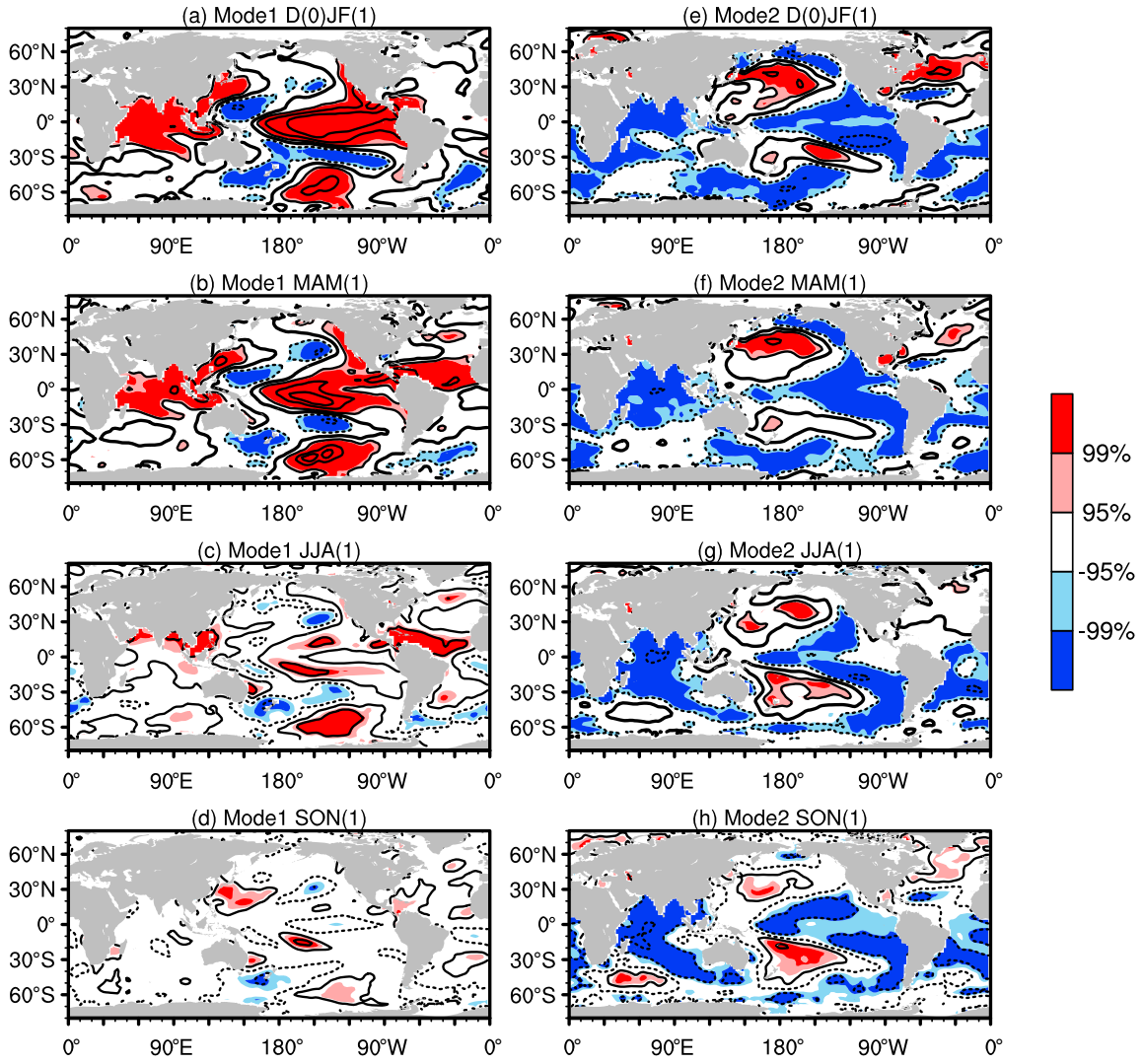

Fig. 9 Correlation between SST and the time series of SST modes in Fig. 7 from the preceding winter to the following autumn.

as the predictor field of the YRI and try to reproduce them using the PLS regression method.

c *Prediction Results Based on the PLS Model*

To predict the YRI time series, we constructed a PLS model. The entire dataset was divided into two parts: a training period used to build the model by using the 47 years of data from 1950 to 1996, which is considered to be sufficient to produce a reliable predictor field, and a test period from 1997 to 2016, which is used to verify the model's predictive ability. The specific process is described in the following. First, the winter SST during the 1949–1995 period and the YRI during the 1950–1996 period were used to obtain the reproduced YRI. By establishing the prediction equation below, we can obtain the regression coefficients β and the residual predictand vector $\mathbf{Y}_{\text{residual}}$. Then we applied the winter SST (1949–1996) and β from the above procedure for the fitted YRI (1950–1997). The last value of $\mathbf{Y}_{\text{residual}}$ is added to the last fitted value to obtain the predicted YRI in 1997. Next, this predicted value was combined with the original YRI (1950–1996) to form a new series (1950–1997) in order to seek the next values for β and $\mathbf{Y}_{\text{residual}}$. In this way, we can finally obtain the retrospective forecast results.

$$\text{YRI} = \beta\,\text{SST} + \mathbf{Y}_{\text{residual}}.$$

We calculated the correlation coefficient between the real YRI index series and the simulation results to obtain the fitting skill and the correlation between the observed YRI and the PLS prediction results to obtain a hindcast skill. Figure 10 shows the simulation and forecast performance of this model. For the training period, the average fitting skill reaches 0.69. Moreover, it generates reasonable skill in the 20-year hindcast period of 1997 to 2016, and the correlation coefficient reaches 0.51. Given that this model uses the preceding winter (DJF) mean SST as the predictor field, this model is able to predict

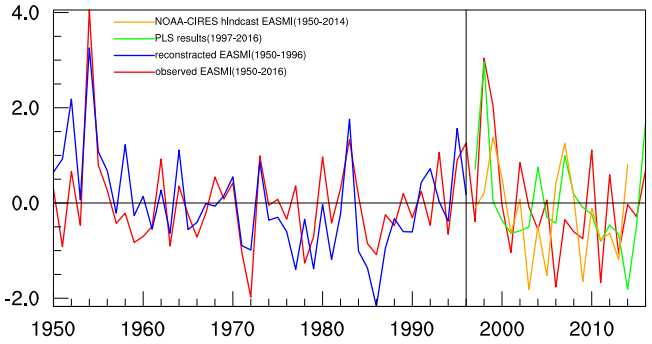

Fig. 10 The PLS prediction model results. The red line is the observed YRI. The blue line is the reconstructed YRI during the training period, 1950–1996. The green line from 1997 to 2016 is the PLS prediction results. The orange line shows the forecast for 1950–1996 and the purple line show the forecast for 1997–2014 from NOAA–CIRES 20CR V2c.

whether a flood or drought will happen three months ahead of the arrival of the flooding season. In order to make a comparison, we calculated the YRI using the NOAA–CIRES 20CR V2c data. These reanalysis data are from a state-of-the-art reanalysis data system and can provide runoff reanalysis data from 1950 to 2014. The correlation coefficient skill is 0.61 for the training period (1950–1996). But for the test period from 1997 to 2016, the correlation coefficient is 0.33. This result is significantly lower than the PLS forecast skill. That is to say, this PLS model shows even better performance than the reanalysis data. This result indicates that this PLS model can be applied to increase prediction accuracy.

In the summers of 1954 and 1998, the YRI was larger than or equal to three, which is perhaps two times as large as in a normal year (Fig. 10). Basin-wide catastrophic floods occurred in the Yangtze River basin in 1954 and 1998 (Yang & Wei, 1999), and our index reflects the runoff changes perfectly. But in 2016, when a very large El Niño occurred in the preceding winter, the PLS produces a higher index than in normal years, which is inconsistent with the actual situation. However, in the past decades, some major hydraulic engineering projects, such as the upstream Three Gorges Reservoir and the South to North Water Diversion Project, were completed that are of critical importance in decreasing the influence of flooding. Some studies have demonstrated that since the Three Gorges Dam started operation in 2003, the runoff from the Yangtze River estuary decreased by 10% (Kuang et al., 2014). This may explain why the PLS prediction can predict the change of flow under the influence of natural factors. In the future, we may make some revisions to the YRI so that the YRI can be used to predict the EASM.

5 The relationship between the YRI and the EASM

There has been much controversy about the way in which to measure the strength of the EASM (Wang et al., 2008). For a long time, many scholars have put forward various definitions of EASM intensity indices based mainly on two factors. Some authors tended to follow the definition of the All-India Rainfall Index, which is a general measure of the strength of the Indian summer monsoon. Most experts use the Meiyu rainfall averaged over a specific Meiyu area to represent the EASM (e.g., Chang, Zhang, & Li, 2000a, 2000b) and Chang et al. focused on average precipitation in a specific Meiyu area. Meiyu is a unique climate feature in the East Asian monsoon region, characterized by persistent rainfall over the middle and lower reaches of the Yangtze River in summer (Ge, Guo, Zheng, & Hao, 2008). However, the JJA mean rainfall anomaly always exhibits large spatial asymmetry over East Asia, partly due to the influence of the mountainous topography; thus, using the Meiyu rainfall to measure EASM variability seems unreliable (Wang et al., 2008). Hence, many investigators began to search for a simple way to define the strength of the EASM, and they tended to use circulation parameters rather than precipitation. However, none of the indices can reflect all the characteristics of the EASM, and so far there has been no consistent definition.

Precipitation and runoff are two critical processes in the hydrological cycle. To the best of our knowledge, many meteorologists have studied precipitation extensively and many hydrologists have studied runoff in detail, but few studies have attempted to address meteorology and hydrology together. However, the flow of the Yangtze River, which is almost parallel to the Meiyu rainbelt, makes the occurrence of floods and droughts in this valley particularly sensitive to interannual variations in rainfall. Considerable year-to-year variability of summer Meiyu rainfall exposes the Yangtze River basin to threats of frequent droughts and floods. Because the correspondence between the long-term annual rainfall and runoff over the Yangtze River basin is good (Xu, Milliman, & Xu, 2010), the newly defined YRI can also capture the features of the summer monsoon precipitation and the interannual variations of the principal components of EASM circulation systems (the rainfall pattern, the SAH, and the SH systems) (Figs 4 and 5), the complex rainfall structure and unreliable precipitation predictions encouraged us to use the runoff index (YRI) to represent EASM intensity.

To test the function of the YRI for measuring the strength of the EASM, we use three EASM indices to validate the performance of the newly defined index. Wang et al. (2008) employed a unified EASM index, which is defined by the difference between U_{850} (22.5°–32.5°N, 110°–140°E) and U_{850} (5°–15°N, 90°–130°E). They also compared many indices and found that the unified EASM index has excellent skill in representing the circulation and precipitation features over East Asia. They also found that the indices proposed by Li and Zeng (2002) and Zhang, Tao, and Chen (2003) can also be good indicators of the EASM (Wang et al., 2008). Li and Zeng's index was defined as a domain-averaged (10°–40°N, 110°–140°E) JJA dynamical normalized seasonality (DNS) at 850 hPa, whereas the index proposed by Zhang et al. (2003) was defined as the difference in JJA 850 hPa anomalous wind between 10°–20°N, 100°–150°E and

TABLE 2. Correlation coefficients between three existing East Asian summer monsoon indices (using NCEP-NCAR data) and the newly defined Yangtze River runoff index (YRI).

	YRI
Iwang	0.49**
Ili	−0.32**
Izhang	−0.37**

Note: Double asterisks (**) represent coefficients that are significant above the 99% confidence level.

25°–35°N, 100°–150°E. Table 2 presents correlation coefficients among the YRI and the existing three EASM indices from 1950 to 2016. Note that the correlation between the YRI and the EASM indices are generally high and reliable. This suggests that the newly defined index (YRI) may be an alternative index for the EASM.

Figure 11 depicts the correlation between EASM indices and the seasonal mean rainfall over East Asia in summer. For comparison purposes, we reverse the sign of Li index and Zhang's (2002) index. The relationship between the EASM and rainfall over East Asia shows remarkable regional features. Comparing with Fig. 4, all these indices can illustrate a negative–positive-negative tripole precipitation pattern in East Asia, but the YRI series has a better correlation than the other indices because its correlation is more significant in the Yangtze River basin as well as in the Yellow River basin. In addition, the YRI suggests that drought years in this basin are linked to a weak EASM, while flooding years are related to a strong EASM, which is consistent with Wang et al.'s (2008) index

In this sense, the prediction of this YRI series can not only be used to predict flood and drought events in the Yangtze River but also to provide a new monitoring tool for measuring the intensity of a monsoon. The proposed YRI emphasizes the importance of the Meiyu-related precipitation in determining the intensity of the EASM.

6 Discussion and conclusions

The Yangtze River plays a crucial role in the economic and social development of China, and the accurate prediction of long-term runoff variation in this basin is important for watershed and flood management. However, frequent floods and droughts in the Yangtze River basin are caused by complex forcing factors, so it is challenging to predict when runoff disasters will happen. The current study focuses on an objective prediction of the variability of runoff during the flooding season.

The YRI proposed in this paper is a meaningful indicator of flood and drought events in the Yangtze River basin. To identify the possible predictors of the runoff variations, a PE PLS regression method was employed to delineate the relationship between the principal SST modes and the YRI. Results show that the SST modes resemble patterns that are highly linked with ENSO or mega-ENSO. It indicates that ENSO or mega-ENSO might be the dominant factor affecting YRI variability.

A PE PLS model is built with the SST modes using a 47-year training period (1950–1996), and forecasts are produced for the 20-year validation period from 1997 to 2016. The PLS model exhibits promising predictive skill with the correlation coefficient between the observation and the prediction reaching 0.51, which is even better than that of the hindcast achieved by the NOAA–CIRES 20CR V2c reanalysis systems. The results have shown that the PLS model has some significant advantages for the long-term prediction of runoff and may provide a new and effective tool for the prediction of drought and flood events in the Yangtze River basin. Meanwhile, the YRI series can also capture the interannual variations of the principal components of the EASM circulation system, thus providing a new monitoring tool for the EASM.

However, it should be mentioned that the variability of the Yangtze catchment hydrology is not only closely linked with

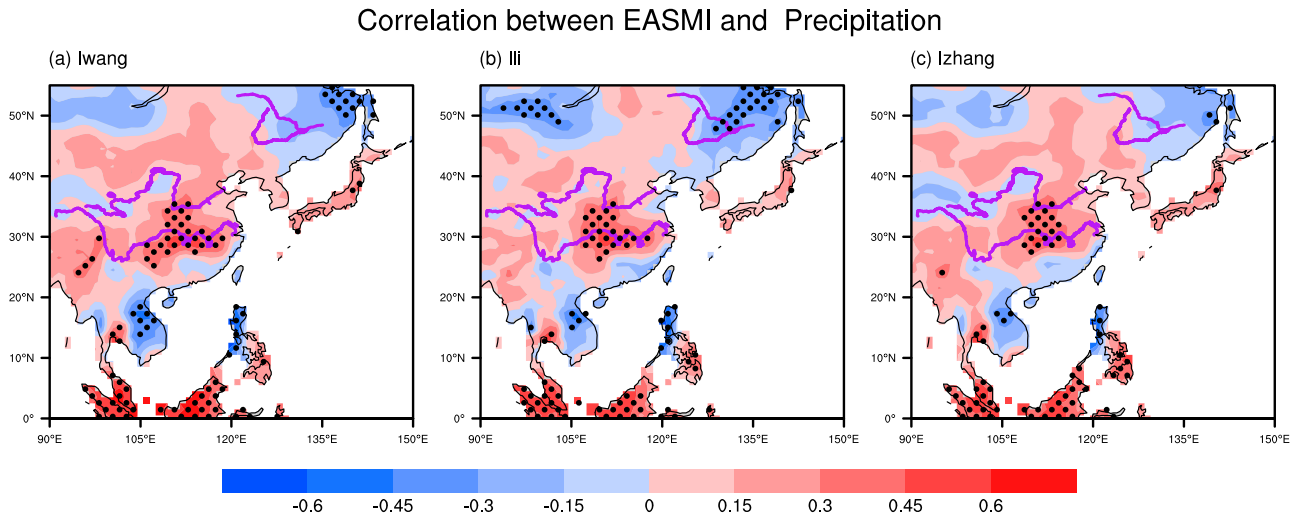

Fig. 11 As in Fig. 4, but for the correlation between the EASM indices and precipitation.

the large-scale atmospheric circulation over East Asia but also with Tibetan Plateau snow cover because the Yangtze originates from the eastern Tibetan Plateau and is fed by the glaciers of the Himalayas (Blender, Zhu, Zhang, & Fraedrich, 2011). Thus, in addition to the SST signal, snow cover or other anomalies may also be predictors of runoff. However, the current findings focus on the importance of ENSO or mega-ENSO in predicting the YRI, and we may conduct a further study in the future to analyze the importance of snow cover.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was jointly supported by the National Natural Science Foundation of China [grant number 41790475], the Ministry of Science and Technology of the People's Republic of China [grant number 2016YFA0601801] [grant number 2015CB453201], and [grant number 2015CB953904], and the National Natural Science Foundation of China [grant number 91637312], [grant number 41575075], and [grant number 91437216].

References

Blender, R., Zhu, X., Zhang, D., & Fraedrich, K. (2011). Yangtze runoff, precipitation, and the East Asian monsoon in a 2800 years climate control simulation. *Quaternary International*, *244*, 194–201.

Chang, C., Zhang, Y., & Li, T. (2000a). Interannual and interdecadal variations of the East Asian summer monsoon and tropical pacific SSTs. Part I: Roles of the subtropical ridge. *Journal of Climate*, *13*, 4310–4325.

Chang, C., Zhang, Y., & Li, T. (2000b). Interannual and interdecadal variation of the East Asian summer monsoon rainfall and tropical SSTs. Part II: Meridional structure of the monsoon. *Journal of Climate*, *13*, 4326–4340.

Chen, M., Xie, P., Janowiak, J. E., & Arkin, P. A. (2002). Global land precipitation: A 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, *3*, 249–266.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., … Brönnimann, S. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*, 1–28.

Dai, F., & Zhu, X. (2010). Application of tendency prediction in trend research on runoff change in the lower reaches of the Yangtze River. *Journal of China Hydrology*, *5*, 42–45.

Ge, Q., Guo, X., Zheng, J., & Hao, Z. (2008). Meiyu in the middle and lower reaches of the Yangtze River since 1736. *Chinese Science Bulletin*, *53*, 107–114.

Guo, S., Guo, J., Hou, Y., Xiong, L., & Hong, X. (2015). Prediction of future runoff change based on Budyko hypothesis in Yangtze River basin. *Advances in Water Science*, *26*, 151–159.

Ju, Q., Hao, Z., Yu, Z., Xu, H., Jiang, W., & Hao, J. (2011). Runoff prediction in the Yangtze River basin based on IPCC AR 4 climate change scenarios. *Advances in Water Science*, *22*, 462–469.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., … Zhu, Y. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*, 437–471.

Kuang, C., Su, P., Gu, J., Chen, W., Zhang, W., & Zhang, Y. (2014). Multi-time scale analysis of runoff at the Yangtze estuary based on the Morlet Wavelet Transform method. *Journal of Mountain Science*, *11*, 1499–1506.

Lai, X., Jiang, J., Liang, Q., & Huang, Q. (2013). Large-scale hydrodynamic modeling of the middle Yangtze River Basin with complex river–lake interactions. *Journal of Hydrology*, *492*, 228–243.

Lee, J., Lee, S., Wang, B., Ha, K., & Jhun, J. G. (2013). Seasonal prediction and predictability of the Asian winter temperature variability. *Climate Dynamics*, *41*, 573–587.

Li, J., & Zeng, Q. (2002). A unified monsoon index. *Geophysical Research Letters*, *29*, 1274. doi:10.1029/2001GL013874

Sampe, T., & Xie, S. (2010). Large-scale dynamics of the Meiyu-Baiu rainband: Environmental forcing by the westerly jet. *Journal of Climate*, *23*, 113–134.

Smith, T., Reynolds, R., Peterson, T., & Lawrimore, J. (2008). Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). *Journal of Climate*, *21*, 2283–2296.

Wang, B., Liu, J., Kim, H., Webster, P., Yim, S., & Xiang, B. (2013). Northern hemisphere summer monsoon intensified by mega-El Niño/Southern Oscillation and Atlantic multidecadal oscillation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 5347–5352.

Wang, B., Wu, Z., Li, J., Liu, J., Chang, C., Ding, Y., & Wu, G. (2008). How to measure the strength of the East Asian summer monsoon. *Journal of Climate*, *21*, 4449–4463.

Wang, B., Xiang, B., Li, J., Webster, P. J., Rajeevan, M. N., Liu, J., & Ha, K. J. (2015). Rethinking Indian monsoon rainfall prediction in the context of recent global warming. *Nature Communications*, *6*, 7154. doi:10.1038/ncomms8154

Wu, Z., & Li, J. (2008). Prediction of the Asian-Australian monsoon interannual variations with the grid-point atmospheric model of IAP LASG (GAMIL). *Advances in Atmospheric Sciences*, *25*, 387–394.

Wu, Z., & Li, J. (2009). Seasonal prediction of the global precipitation annual modes with the grid-point atmospheric model of IAP LASG. *Journal of Meteorological Research*, *23*, 428–437.

Wu, Z., Li, J., Jiang, Z., & He, J. (2011). Predictable climate dynamics of abnormal East Asian winter monsoon: Once-in-a-century snowstorms in 2007/2008 winter. *Climate Dynamics*, *37*, 1661–1669.

Wu, Z., Lin, H., Li, Y., & Tang, Y. (2013). Seasonal prediction of killing-frost frequency in South-Central Canada during the cool/overwintering-crop growing season. *Journal of Applied Meteorology and Climatology*, *52*, 102–113.

Wu, Z., Wang, B., Li, J., & Jin, F. (2009). An empirical seasonal prediction model of the East Asian summer monsoon using ENSO and NAO. *Journal of Geophysical Research: Atmospheres*, *114*, D18120. doi:10.1029 /2009JD011733

Wu, Z., & Yu, L. (2015). Seasonal prediction of the East Asian summer monsoon with a partial-least square model. *Climate Dynamics*, *46*, 3067–3078.

Xu, J., He, Q., Liu, H., & Chen, J. (2008). Preliminary analysis of characteristics of the exceptional low discharge and its cause over the Yangtze River, 2006 [in Chinese]. *Resources and Environment in the Yangtze Basin*, *5*, 716–722.

Xu, K., Milliman, J. D., & Xu, H. (2010). Temporal trend of precipitation and runoff in major Chinese rivers since 1951. *Global and Planetary Change*, *73*, 219–232.

Yang, Y., & Wei, Z. (1999). Comparison and consideration of the Yangtze River flood in 1998 and 1954 [in Chinese]. *Meteorological Science and Technology*, *1*, 16–19.

Zhang, Q., Tao, S., & Chen, L. (2003). The interannual variability of East Asian summer monsoon indices and its association with the pattern of general circulation over East Asia [in Chinese]. *Acta Meteorologica Sinica*, *61*, 559–568.

Zhao, J., Li, J., Dai, Z., Zhang, H., & Yan, H. (2010). Key role of the lakes in runoff supplement in the mid-lower reaches of the Yangtze River during typical drought years. In *2010 International Conference on Digital Manufacturing and Automation (ICDMA)*, 18-20 December 2010, Changsha, China, 61, 874–880.

Zhao, Y., Xu, X., Zhao, T., Xu, H., Mao, F., Sun, H., & Wang, Y. (2016). Extreme precipitation events in East China and associated moisture transport pathways. *Science China Earth Science*, *59*, 1854–1872.