Improvement of decadal predictions of monthly extreme Mei-yu rainfall via a causality guided approach

ENVIRONMENTAL RESEARCH
CLIMATE

PURPOSE-LED PUBLISHING™

Part of the ENVIRONMENTAL RESEARCH SERIES

LETTER • **OPEN ACCESS**

Improvement of decadal predictions of monthly extreme Mei-yu rainfall via a causality guided approach

To cite this article: Kelvin S Ng *et al* 2024 *Environ. Res.: Climate* **3** 041001

ENVIRONMENTAL RESEARCH
CLIMATE

CrossMark

OPEN ACCESS

RECEIVED
2 July 2024

ACCEPTED FOR PUBLICATION
22 July 2024

PUBLISHED
14 August 2024

LETTER

Improvement of decadal predictions of monthly extreme Mei-yu rainfall via a causality guided approach

Kelvin S Ng[1,*], **Gregor C Leckebusch**[1] **and Kevin I Hodges**[2]

[1] School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, United Kingdom
[2] Department of Meteorology, National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: k.s.ng@bham.ac.uk

Keywords: extreme rainfall, Mei-yu front, causality-guided approach, decadal prediction system

Supplementary material for this article is available online

Abstract

While the improved performance of climate prediction systems has allowed better predictions of the East Asian Summer Monsoon rainfall to be made, the ability to predict extreme Mei-yu rainfall (MYR) remains a challenge. Given that large scale climate modes (LSCMs) tend to be better predicted by climate prediction systems than local extremes, one useful approach is to employ causality-guided statistical models (CGSMs), which link known LSCMs to improve MYR prediction. However, previous work suggests that CGSMs trained with data from 1979–2018 might struggle to model MYR in the pre-1978 period. One hypothesis is that this is due to potential changes in causal processes, which modulate MYR in different phases of the multidecadal variability, such as the Pacific decadal oscillation (PDO). In this study, we explore this hypothesis by constructing CGSMs for different PDO phases, which reflect the different phases of specific causal process, and examine the difference in quality as well as with respect to difference drivers and thus causal links between CGSMs of different PDO phases as well as the non-PDO phase specific CGSMs. Our results show that the set of predictors of CGSMs is PDO phase specific. Furthermore, the performance of PDO phase specific CGSMs are better than the non-PDO phase specific CGSMs. To demonstrate the added value of CGSMs, the PDO phase specific versions are applied to the latest UK Met Office decadal prediction system, DePreSys4, and it is shown that the root-mean squared errors of MYR prediction based on PDO phase specific CGSMs is consistently smaller than the MYR predicted based on the direct DePreSys4 extreme rainfall simulations. We conclude that the use of a causality approach improves the prediction of extreme precipitation based solely on known LSCMs because of the change in the main drivers of extreme rainfall during different PDO-phases.

1. Introduction

Increasing the capability of predicting and representing extreme rainfall over East Asia, the home of more than 1.6 billion people (United Nations 2022), is of importance as extreme rainfall poses a significant threat to societies via the subsequent regional flood hazards. Some unprecedented extreme rainfall events have occurred over China during the East Asian Summer Monsoon (EASM) season in the past few years (e.g. Ding *et al* 2021, Zhou *et al* 2022, Wu *et al* 2023). These extreme rainfall events have induced significant socioeconomic impacts on China (e.g. Zhou *et al* 2022). Recent advances in near-term (seasonal and decadal) climate prediction systems have been shown to provide useful predictions of rainfall over China during EASM on seasonal and monthly timescales from direct model outputs (Li *et al* 2016, Martin *et al* 2020) as well as via an EASM index (e.g. Wang and Fan 1999) to infer rainfall (Bett *et al* 2020, 2021, 2023, Martin *et al* 2020). The Mei-yu rainfall, which is rainfall triggered by the Mei-yu front (MYF), is responsible for 45% of total summer rainfall in the middle/lower Yangtze River Valley (Ding and Chan 2005). Thus, to further improve the ability of near-term climate predictions of extreme precipitation over China during EASM, it is necessary to enhance our ability to capture the major contributors to EASM rainfall over China—the Mei-yu rainfall (e.g. Bett *et al* 2021).

Extreme Mei-yu rainfall (hereinafter MYR) is known to be influenced by various large-scale climate modes (LSCMs), such as the South Asian High (Liu and Ding 2008, Ning *et al* 2017), the Western North Pacific Subtropical High (Zhou and Wang 2006, Ninomiya and Shibagaki 2007, Liu and Ding 2008, Sampe and Xie 2010, Ding *et al* 2021), and the El Nino–Southern Oscillation (Wu *et al* 2003, Wang *et al* 2009, Ye and Lu 2011). Since climate models can produce LSCMs better than extreme rainfall (Flato *et al* 2013), a way to improve MYR simulations is to make use of these relevant LSCMs as (physical) transmitters of information to construct statistical models (Ng *et al* 2022). A similar approach, but based on traditional correlation estimates, has been applied in seasonal forecast of monthly mean and seasonal mean Mei-yu rainfall (Bett *et al* 2020, 2021, Martin *et al* 2020) as well as in extended seasonal forecast (Bett *et al* 2023).

However, traditional correlation approaches to identify useful predictors to construct statistical models suffer from the inability to identify robust and comprehensive relationships between MYR and LSCMs due to the complexity of the EASM system. To overcome this issue, Ng *et al* (2022) introduced statistical models constructed using a causality algorithm (Tigramite v4.2; Runge *et al* 2019, Runge 2020). Many studies have demonstrated the usefulness of the causality approach in statistical model construction to model different types of atmospheric phenomena, such as extreme stratospheric polar vortex states (Kretschmer *et al* 2017), regional Indian summer monsoon rainfall (Di Capua *et al* 2019) and MYR (Ng *et al* 2022). The main advantage of using the causality approach, as opposed to correlation approaches, is that causality-based models do not suffer from overfitting due to non-causal relationships between predictors and predictand (Kretschmer *et al* 2017). This also implies that the causality-based models are transferable, i.e. can be applied to other data, as causality-based models can capture the underlying physical (causal) relations between predictors and predictands rather than the simple association between predictors and predictands.

Using the principle of causality, Ng *et al* (2022) constructed causality-guided statistical models (CGSMs) to model the MYR using known LSCMs in the period of 1979–2018. They demonstrated that CGSM can capture important observed characteristics of the MYR, such as the sub-monthly variability, as well as physical significance based on the spatial coherency of the choice of predictors. However, Ng *et al* (2022) noticed that CGSMs struggled to model the MYR in the period of 1961–1978. They hypothesized that the lack of representation of low frequency oceanic variability could be the reason. Indeed, several studies have documented decadal and interdecadal variability of the EASM precipitation pattern, which are linked to the Pacific decadal oscillation (PDO) and the Atlantic multidecadal oscillation (AMO) (e.g. Ding *et al* 2008, 2018, 2020). This raises the scientific question: Is the MYR driven by different sets of causal predictors in different phases of the multi-decadal oscillation, i.e. is the stationarity condition satisfied? If the stationarity condition is not satisfied, statistical models should be constructed based on the respective phase of multi-decadal oscillations. This has significant consequences when CGSM would be applied to climate prediction data where decadal oscillations could play a major role in modulating MYR.

The objective of this study are:

(i) To explore the stationarity of the causal predictors in different phases of a decadal/multidecadal variability mode.
(ii) To demonstrate a general workflow for the application of causality approach that would be useful in improving MYR representation in climate predictions.

In this study, we first introduce an improved approach, CGSM2 (see section 3.1), to construct CGSMs by utilizing spatial coherency of predictors. Then, we use CGSM2 with the PDO phase specific set of causal predictors to show the validity of the concept of different causality models for different PDO phases in comparison to stationary causality models. Due to data availability, we can only investigate the relationship between one decadal and interdecadal variability and MYR. PDO is chosen due to its highly non-linear relationship with the EASM precipitation pattern (Ding *et al* 2018), which is of the particular interest from the application perspective. Furthermore, this study aims to test the validity of the concept. To demonstrate the added value of the proposed approach, the performance of the PDO phase specific CGSM2 in modelling MYR is compared with the direct model outputs of the latest UK Met Office Decadal Prediction System version 4, DePreSys4 (Scaife *et al* 2022). The paper is organised as follows: sections 2 and 3 describe the data and methodologies used in this study. Main results are displayed in section 4. Discussion and conclusions are presented in section 5.

Table 1. List of the LSCMs considered in the construction of CGSM. All with lead time from 1 month up to 11 months.

LSCM	Definition/References
Dipole mode index (DMI)	As in Saji *et al* (1999), Black *et al* (2003)
Indian monsoon index by wang and fan (IMI-WF)	As in Wang and Fan (1999)
Indian monsoon index by webster and yang (IMI-WY)	As in Webster and Yang (1992)
ENSO (Nino 3.4)	As in Trenberth (1997)
Pacific Japan pattern (PJ)	As in Nitta (1987), Wakabayashi and Kawamura (2004), Choi *et al* (2010), Kim *et al* (2012), Li *et al* (2013)
South Asian High Area Index (SAHI-Area)	As in Ning *et al* (2017)
South Asian High Northwest Displacement Index (SAHI-NW)	As in Ning *et al* (2017)
Silk Road pattern principal component 1 (SRP-PC1)	As in Li *et al* (2020)
Silk Road pattern principal component 2 (SRP-PC2)	As in Li *et al* (2020)
Sea surface temperature anomaly of Arabian Sea (SSTA-AS)	Mean SST anomaly in the region 10–25° N, 60–75° E
Sea surface temperature anomaly of Bay of Bengal (SSTA-BoB)	Mean SST anomaly in the region 10–23° N, 80–100° E
Sea surface temperature anomaly of East China Sea (SSTA-ECS)	Mean SST anomaly in the region 25–33° N, 120–130° E
Sea surface temperature anomaly of South China Sea (SSTA-SCS)	Mean SST anomaly in the region 10–23° N, 105–120° E
Western North Pacific monsoon index (WNPMI)	As in Wang and Fan (1999), Wang *et al* (2001), Wang *et al* (2008)
Western North Pacific Subtropical High North Index (WNPSH-North)	As in Lu (2002)
Western North Pacific Subtropical High West Index (WNPSH-West)	As in Lu (2002)

2. Data

Observed precipitation data over China from 1961 to 2018 is obtained from CN05.1 (Wu and Gao 2013). CN05.1 is a high resolution (0.25° × 0.25°) gridded daily data set constructed by interpolating data from more than 2400 observation stations in China using the 'anomaly approach' (Xu *et al* 2009, Wu and Gao 2013).

The European Centre for Medium-Range Weather Forecasts fifth generation reanalysis data (ERA5, Hersbach *et al* 2020) and ERA5 back extension (preliminary version) (Bell *et al* 2020a, 2020b), with spatial resolution of 0.25° × 0.25°, were used to calculate indices of known LSCMs (table 1), MYF and tropical cyclone detection (see section 3.2).

Observed NCEI PDO indices are obtained from National Oceanic and Atmospheric Administration (NOAA 2022). The indices are derived based on the extended reconstruction sea surface temperature (SST) dataset version 5 (ERSSTv5, Huang *et al* 2017). Detailed description of the related methodology is available at www.ncei.noaa.gov/access/monitoring/pdo/ (last accessed 13 July 2023).

To demonstrate the added value of the PDO phase specific CGSM2, hindcast outputs of DePreSys4 (Scaife *et al* 2022) have been used. The setup of DePreSys4 is based on the HadGEM3-GC3.1-MM historical simulations suite (Williams *et al* 2018). The hindcast outputs were generated following the decadal climate prediction project of CMIP6 component A protocol (Boer *et al* 2016). DePreSys4 has 10 ensemble members, and hindcasts were initialised on the 1st of November for every year in the period of 1960–2018 with 10 years hindcast period. In this study, DePreSys4 hindcast outputs of lead year 2–10, which are initialised from 1960 to 2009, are used. This study period ensures all hindcast years can be evaluated against observations and the number of hindcast for each lead year are identical. Hindcasts of lead year 1 are not used as the hindcast data do not cover the period where causal predictors can be identified (see section 3.1). The total number of model years evaluated is 4,410. HadGEM3-GC3.1-MM historical simulations (Andrews *et al* 2020) are used to derive the PDO patterns of the HadGEM3-GC3.1 system for the PDO index (see section 3.3).

3. Methods

3.1. CGSM with consideration of spatial coherency—CGSM2

CGSM2 aims to use the information of monthly LSCMs, i.e. predictors, to infer total monthly MYR, i.e. predictand. For each grid box of a given month of interest, a CGSM2 is constructed if there are at least 10 non-zero MYR entries. CGSM2 contains three main steps: (i) causal predictor selection for each grid box; (ii) evaluation of the spatial coherency of the causal predictors; and (iii) model construction for each grid box. In comparison to CGSM (Ng *et al* 2022), CGSM2 utilises the notion of spatial coherency of causal predictors, which further constrains the choice of causal predictor and reduces the possibility that a predictor is selected due to purely statistical optimisation. The schematic workflow of CGSM2 can be found in figure S1. A brief description of CGSM2 is as follows.

3.1.1. Causal predictor selection for each grid box

To capture known and unknown relationships between known LSCMs and MYR using causality approach, a pool of potential predictors is constructed using the indices of known LSCM (table 1) with a lead time of up to 11 months prior to the month of interest. The total number of potential predictors in the pool is 176. The use of potential predictors with various lead times aims to compensate for the incomplete set of potential predictors. LSCMs of large lead time could increase the likelihood of capturing 'missing LSCMs' because they would act as proxies for hidden processes.

The correlation between the potential predictors and the predictand is calculated. Conditional dependency between the potential predictors, which are significantly correlated (p-value < 0.1; as in Ng *et al* (2022)) with the predictand, and the predictand is evaluated using the modified Peter-Clark algorithm (Tigramite v4.2; Runge *et al* 2019, Runge 2020). Causal predictors are identified if the predictand is conditionally dependent on them. This forms a preliminary set of causal predictors.

3.1.2. Evaluation of the spatial coherency

Ng *et al* (2022) showed that the grid boxes of a given CGSM predictor would form spatially coherent clusters, indicating these predictors have physical meaning and significance. They suggested to use those spatially coherent clusters to further constrain the choice of causal predictors and reduce the possibility that a predictor is selected due to purely statistical optimisation. CGSM2 utilises the notion of spatial coherency in the predictor selection step by incorporating the density-based spatial clustering (DBSCAN; Ester *et al* 1996). For a given causal predictor X, the spatial positions of the grid boxes that have causal predictor X in the preliminary set of causal predictors are first identified. DBSCAN is then applied to the spatial positions of these grid boxes, and subsequently, spatial clusters are identified. Figure 1 shows a schematic example of the DBSCAN output.

Unlike other commonly used clustering algorithms, such as K-means clustering, DBSCAN does not require a pre-defined number of clusters and the clusters identified by DBSCAN are not constrained by a specific shape or relative size to other clusters. This provides the necessary flexibility to capture clusters with potentially erratic shapes and uneven sizes. A similar approach has been successfully applied to objectively identify tropical cyclone cloud clusters from satellite images in Ng *et al* (2020). Clusters smaller than the minimum cluster size, are labelled as noise. Predictors of the noise clusters are removed from the preliminary set of causal predictors and are not used in model construction. The minimum cluster size is chosen to be 16 grid boxes (equivalent to a $1^\circ \times 1^\circ$ grid box). While the threshold of 16 grid boxes is an arbitrary choice, this threshold is effective in removing isolated points and maintaining the performance of CGSM2.

3.1.3. Model construction for each grid box

For each grid box, a linear model is constructed between the set of causal predictors identified in section 3.1.2 and predictand using multiple linear regression. Depending on the month of interest and period of consideration, the total number of models constructed ranges from 3142 to 6488. Typically two to five causal predictors are used in model construction (table S1).

3.2. Identification of extreme Mei-yu precipitation

MYR is defined as the extreme rainfall above the 95th percentile of the local climatological daily rainfall within 500 km north and south of the MYF, excluding precipitation caused by tropical cyclones. Position of MYFs were detected by a scheme described in Ng *et al* (2022). The MYF detection scheme locates the daily position of MYF by using the minimum of the product of the meridional gradient of daily equivalent potential temperature at 850 hPa and specific humidity at 850 hPa. MYFs in ERA5 are identified following the aforementioned description. For DePreSys4, as daily data of specific humidity and temperature at 850 hPa level are not available, monthly data were used instead for monthly MYF identification. The MYF

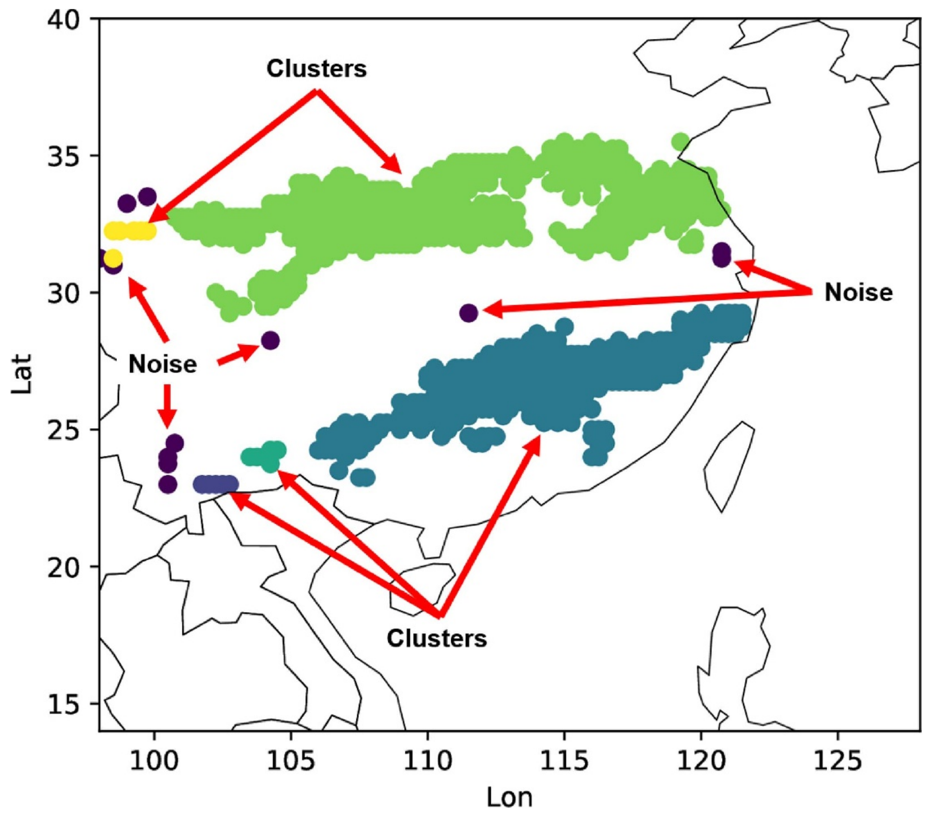

Figure 1. A schematic example of a DBSCAN output. Each dot represents a grid box with a given predictor chosen by the CGSM procedure. Black indicates noise whereas other colours indicate different clusters identified by DBSCAN.

climatology of DePreSys4 constructed using monthly data is similar to the MYF climatology of HadGEM3-GC3.1-MM historical simulations constructed using daily data (figure S2). Tropical cyclone-related rainfall is defined as any rainfall within a 500 km radius of the centre of a tropical cyclone, where the location of the tropical cyclone is identified by the TRACK algorithm (Hodges *et al* 2017).

3.3. Determination of PDO phase

Since CGSM2 can be constructed using indices of known LSCM with lead time up to 11 months prior to the month of interest, the PDO phase is determined by the mean value of PDO indices over the 12 month period ending with the month of interest. This ensures that the PDO phase is not biased by any high-frequency variability, making the approach self-contained. The main findings of the current study are not sensitive to a longer calculation period for determining the PDO phase. For the construction of the CGSM2, the observed NCEI PDO indices (c.f. Section 2) were used. The ratio between the number of years with PDO+ and the number of years with PDO- is roughly 4:6.

For DePreSys4, the PDO index is calculated based on the approach of Mantua *et al* (1997), and Boer and Sospedra-Alfonso (2019): First the PDO pattern, i.e. the leading empirical orthogonal function of the detrended monthly SST anomalies in the North Pacific Ocean poleward of 20° N with global monthly mean SSTs subtracted, from HadGEM3-GC3.1-MM historical simulation is obtained. Then projecting the PDO pattern of the HadGEM3-GC3.1-MM historical simulations onto the detrended DePreSys4 SST with seasonal cycle removed. This ensures the derived PDO pattern is purely from the long-term intrinsic variability of the HadGEM3-GC3.1 system and not influenced by initial conditions and boundary conditions for each initialisation of DePreSys4.

4. Results

The PDO phase specific CGSM2s are constructed by separating training data into PDO+ and PDO- (see section 3.3), and the corresponding models are referred to as PDO+ CGSM2 and PDO- CGSM2, respectively. The CGSM2 constructed using the full period (1961–2018) is referred to as the full period CGSM2. Figure 2 shows the top 10 most frequently selected predictors (table S2) per grid cell chosen in PDO+ CGSM2 and PDO- CGSM2 for each month in the EASM season. For each month, the predictor clusters (i.e. the colour patches; see table S2 for the corresponding predictor of each colour), show different spatial patterns and

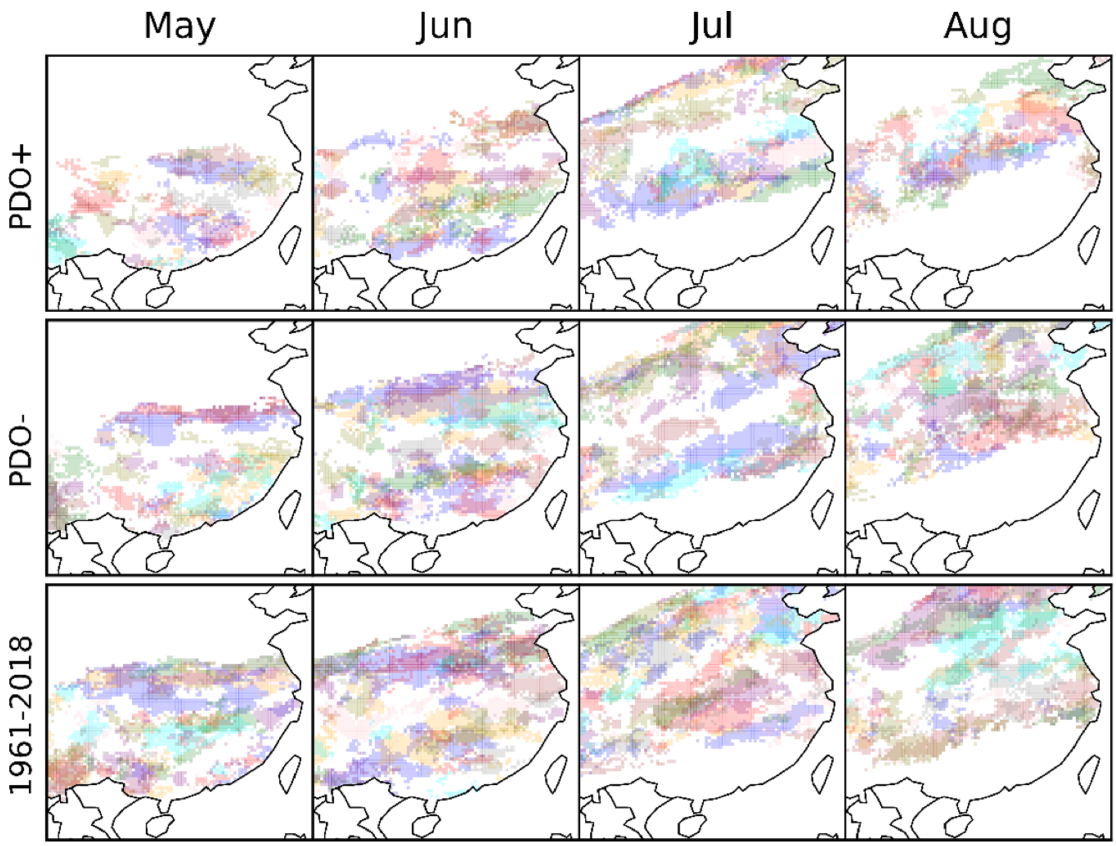

Figure 2. Map of top 10 popular predictors chosen using CGSM2 constructed with data in the period of PDO+ (top), PDO- (middle), 1961–2018 (bottom) for (left to right) May, June, July, and August. The predictor that is represented by any colour in a given panel is independent to other panels in this figure. The list of predictors that are represented by each colour in each panel are stated in table S2.

organisation. This indicates there are systematic changes in the set of causal predictors in different PDO phases. A similar observation can be made when comparing the PDO phase specific maps of predictors with the map of predictors derived using the full period (1961–2018) (figure 2). This shows that the full period CGSM2 attempts to identify causal predictors that would satisfy both PDO phases but not a particular PDO phase. It should be noted that the detailed investigation on the physical linkage between individual causal predictor and MYR is beyond of the scope of this study (further discussion can be found in section 5).

The performance of PDO+ CGSM2, PDO- CGSM2, and full period CGSM2 is shown in figure 3. Across all months in the EASM season, full period CGSM2 (figure 3 bottom row) has the lowest performance, with the overall mean Pearson correlation coefficient (r) of 0.59–0.63 (table 2). CGSM2 constructed using PDO phase specific data have higher performance in comparison to full period CGSM2 (significant at 0.0001 level) with the overall mean r of 0.75–0.79 and 0.68–0.74, for PDO+ CGSM2 and PDO- CGSM2, respectively (table 2). The results are consistent across all months of interest, and they have been validated using 10 000 times repeated five-fold cross validation. This demonstrates that the CGSM2 constructed based on PDO phase specific data can better capture the spatiotemporal variability of the MYR. As the set of causal predictors, which modulates MYR, in different PDO phases is different, this implies the underlying physical mechanisms that modulate the MYR in different PDO phases are different. An interesting observation is that PDO+ CGSM2 has higher performance than PDO- CGSM2 (figure S3). A possible explanation is that since the ratio between number of years with PDO+ and number of years with PDO- is roughly 4:6, i.e. the data set of PDO- is longer than the data set of PDO+. Since other low frequency variability, e.g. AMO could also be modulating factors of MYR, longer datasets may be more likely to be less stationary and consequently have lower performance. Another possible explanation is that the current set of known LSCMs does not include all the necessary LSCMs, related to the MYR modulation during PDO-. This requires further investigation.

To demonstrate the added value of the CGSM2 approach in modelling the MYR, the PDO phase specific CGSM2 are applied to the indices of known LSCMs of DePreSys4, and compared with the direct model outputs of MYR from DePreSys4. In order to correctly assess the added value of CGSM2, the following evaluation criteria were applied: (1) For each forecasted year of a member, the predicted MYR by a DePreSys4 member was only compared to the predicted MYR by PDO phase specific CGSM2, if the PDO

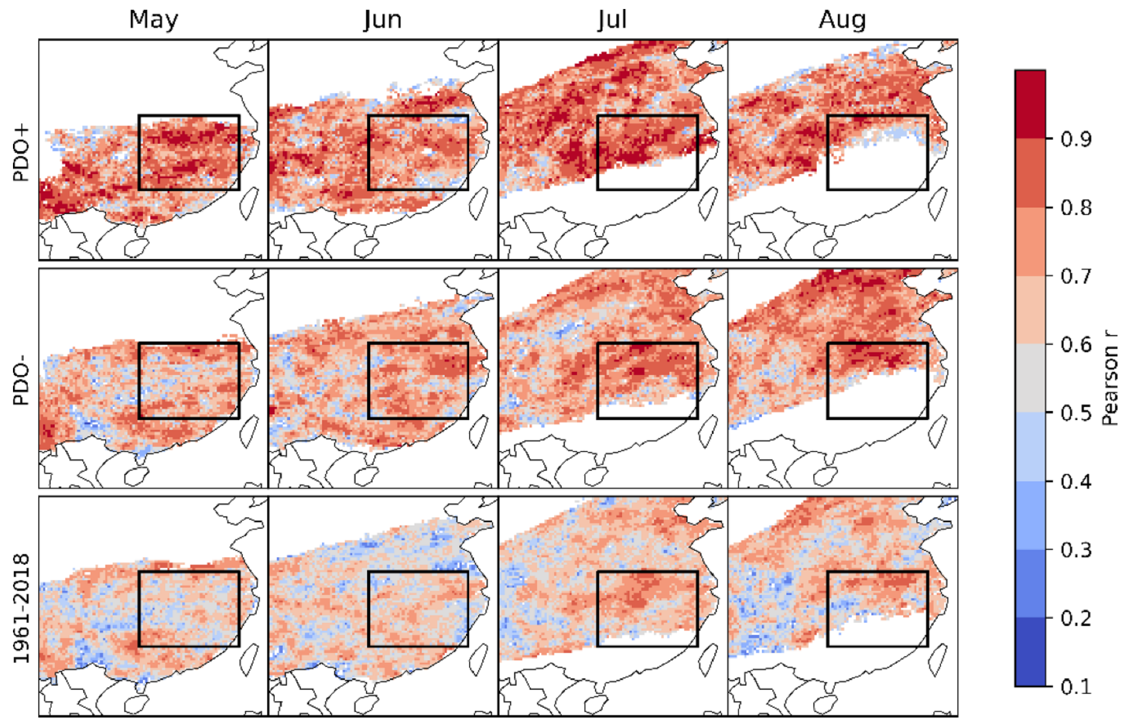

Figure 3. Performance of monthly CGSM2 (Pearson r) constructed using data in the period of PDO+ (top), PDO- (middle), 1961–2018 (bottom), for (left to right) May, June, July, and August with CN05.1 as reference. The black box indicates the region middle/lower Yangtze River Valley.

Table 2. Mean and standard deviation (in brackets) of regional performance, quantified by Pearson correlation coefficient (r), of CGSM2 (figure 3) of different months.

	May	June	July	August
PDO+	0.77(0.13)	0.75(0.13)	0.79(0.12)	0.75(0.13)
PDO-	0.68(0.13)	0.68(0.13)	0.71(0.12)	0.74(0.12)
1961–2018	0.62(0.11)	0.59(0.11)	0.63(0.11)	0.63(0.12)

phase was correctly predicted by that member of DePreSys4. The number of correctly predicted PDO phase for each lead year by the DePreSys4 is shown in figure S4; (2) The performance of the MYR prediction of lead year one was not evaluated. This is because CGSM2, following the basic principle approach of Ng *et al* (2022), considers LSCMs with large lead time (up to 11 months ahead of the month of interest). As discussed in Ng *et al* (2022), there could exist hidden (physical) processes, which are not documented in literature but are important in MYR modulation, the use of large lead time of known LSCMs would assist the models to capture hidden process indirectly. Since the hindcast outputs of DePreSys4 were initialised in November of each year, it does not cover the necessary range of data that is required as input of CGSM2 for MYR prediction of lead year one. (3) The evaluation is to compare the prediction to observations, i.e. CN05.1, of the regional mean MYR over the middle/lower Yangtze River Valley as indicated by the black boxes in figure 3, as this region experiences significant amount of MYR throughout the EASM season (Ding and Chan 2005), posing significant flood risk. Figure 4 shows the root-mean-squared-error (RMSE) of the predictions for each month. MYR predicted using PDO phase specific CGSM2 has a lower RMSE, in comparison to the DePreSys4 direct predictions for both PDO phases for all the lead years, i.e. lead years 2–10, investigated. The performance of both DePreSys4 and PDO phase specific CGSM2 is relatively constant for all lead years of interest. This again confirms the usefulness of the CGSM2 approach.

5. Discussion and conclusions

The performance of CGSM2 constructed using data in the period of 1961–2018 is significantly lower than PDO phase specific CGSM2 (figure 3, table 2). Furthermore, it can be shown that the overall performance of CGSM2 constructed using data in the period of 1979–2018 is statistical significantly higher than the CGSM2 constructed using data in the period of 1961–2018 across all months of interest (figure S5). This highlights a potential issue with using long timeseries for statistical model building—violation of the stationarity

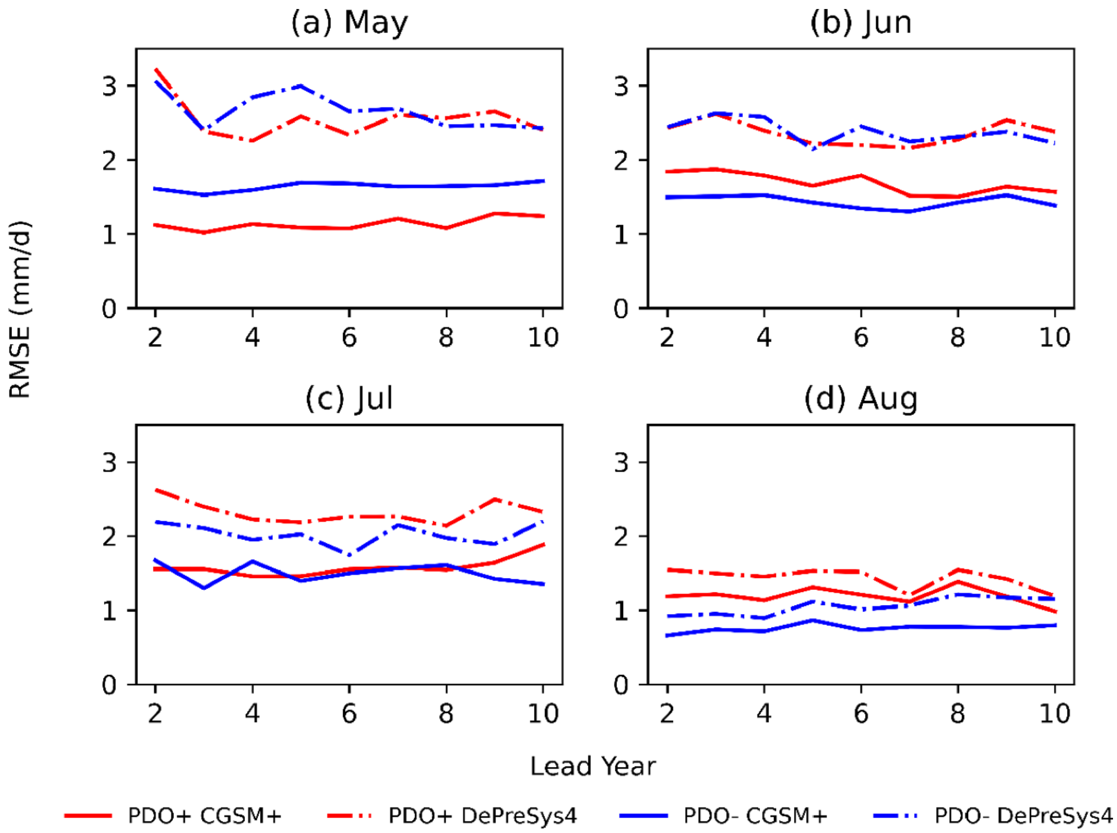

Figure 4. RMSE of monthly mean MYR (mm/d) in the middle/lower Yangtze River Valley predicted by PDO phase specific CGSM2 (solid lines) and DePreSys4 direct extreme precipitation simulation (broken lines) in PDO+ (red) and PDO- (blue) phases for (a) May, (b) June, (c) July, and (d) August for lead years 2–10.

assumption. Based on this observation, future analysis, in addition to different phases of PDO, could also be performed based on different AMO phases as AMO is also known to be a modulating factor of the EASM rainfall (Ding *et al* 2018). This is beyond the scope of the current study as well as we are limited by the number of available reliable observations. One idea to increase the number of observations is to make use of century long observations and reanalysis. However, the quality of the earlier period of century long datasets tend to suffer from significant uncertainty due to changes in the number of observations and their distributions in the early part of the century. Consequently, constructing CGSM2 based on century long data could be unreliable. An alternative approach to overcome this issue is to make use of a so-called Osinski–Thompson approach, also known as the UNSEEN approach (Osinski *et al* 2016, Thompson *et al* 2017), as suggested by Ng *et al* (2022). The Osinski–Thompson approach aims to increase the number of 'observations' by using ensemble outputs of state-of-the-art climate models. This approach has been applied to various studies of extreme events (e.g. Angus and Leckebusch 2020, Ng and Leckebusch 2021). While simulated rainfall in climate models tends have bias due to sub-grid scale parameterisation schemes, this could be addressed using bias correction technique. As demonstrated in previous section, CGSM2 can reduce RMSE of MYR prediction in DePreSys4 using known LSCMs (figure 4). This shows that in addition to improvement in sub-grid scale parameterisation schemes, improving extreme rainfall prediction in climate models could also be done by increasing the skills in predicting known LSCMs.

While the current approach may not have accounted for all relevant large-scale environmental factors known to influence MYR, such as the upper level westerly jet over East Asia (Kong and Chiang 2020, Zhou *et al* 2021, He *et al* 2023), statistically, using LSCMs with large lead time as proxies appears to indirectly capture these factors. However, the linkage between these proxies and the relevant large-scale environmental factors and MYR remains an open question. Further investigations are required to explore these factors. Nevertheless, our approach presents an opportunity to study the EASM and its associated extremes using data-driven approaches.

In conclusion, a new version of the CGSM, CGSM2, which incorporated the notion of spatial coherency of predictor selection, has been developed. Using CGSM2, we have shown that there exists different causal predictor set in different PDO phases. The CGSM2 constructed using PDO phase specific data has better performance than CGSM2 that are constructed using long-term data (1961–2018) with all PDO phases together. This is linked to the fact that using long-term data to construct CGSM2 does not necessarily satisfy the stationarity condition that is required for construction of statistical model. This is because there obviously exist different sets of causal predictors that modulate MYR in different PDO phases. To demonstrate the added value of CGSM2, the PDO phase specific CGSM2 was applied to DePreSys4 and we could show that PDO phase specific CGSM2 performs consistently better than the MYR predicted based on direct DePreSys4 output. Consequently, it provides evidence that causality approach can be useful in improving climate prediction. We have thus demonstrated the advantage of combining machine learning methods, classical statistical approaches and state-of-the-art dynamical model to produce a better representation of extreme rainfall in climate predictions.

Data availability statement

ERA5 and ERA5-BE were obtained from the Copernicus Climate Data Store (CDS). PDO indices are obtained from NCEI, available at www.ncei.noaa.gov/access/monitoring/pdo/ (last accessed 13 July 2023). Historical simulation of CMIP6 outputs and Monthly DePreSys4 hindcast outputs were obtained from the Centre for Environmental Data Analysis (CEDA) via JASMIN. Sub-daily DePreSys4 hindcast outputs were provided by Dr Leon Hermanson at the UK Met Office. CN05.1 was provided by Dr Jia Wu at National Climate Center, China Meteorological Administration.

The data that support the findings of this study are openly available at the following URL/DOI: 10.25500/edata.bham.00001150.

Acknowledgments

This work was supported by the UK-China Research and Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. The calculations described in this paper were performed using the Blue-BEAR HPC service at the University of Birmingham and JASMIN, the collaborative data analysis facility in the UK. The authors thank Dr Leon Hermanson at the UK Met Office for providing high temporal resolution DePreSys4 data for TC tracking and Dr Jia Wu at National Climate Center, China Meteorological Administration for providing CN05.1.

ORCID iDs

Kelvin S Ng https://orcid.org/0000-0002-1464-1701
Gregor C Leckebusch https://orcid.org/0000-0001-9242-7682
Kevin I Hodges https://orcid.org/0000-0003-0894-229X

References

Andrews M B *et al* 2020 Historical simulations with HadGEM3-GC3.1 for CMIP6 *J. Adv. Model. Earth Syst.* **12** e2019MS001995

Angus M and Leckebusch G C 2020 On the dependency of atlantic hurricane and european windstorm hazards *Geophys. Res. Lett.* **47** e2020GL090446

Bell B *et al* 2020a *ERA5 hourly data on pressure levels from 1950 to 1978 (preliminary version) Copernicus Climate Change Service (C3S) Climate Data Store (CDS): Copernicus Climate Change Service (C3S) Climate Data Store* (CDS) (available at: https://cds.climate.copernicus-climate.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels-preliminary-back-extension?tab=overview) (Accessed 15 June 2021)

Bell B *et al* 2020b *ERA5 hourly data on single levels from 1950 to 1978 (preliminary version) Copernicus Climate Change Service (C3S) Climate Data Store (CDS): Copernicus Climate Change Service (C3S) Climate Data Store* (CDS) (available at: https://cds.climate.copernicus-climate.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-preliminary-back-extension?tab=overview) (Accessed 15 June 2021)

Bett P E *et al* 2020 Seasonal rainfall forecasts for the Yangtze River Basin of China in summer 2019 from an improved climate service *J. Meteorol. Res.* **34** 904–16

Bett P E, Dunstone N, Golding N, Smith D and Li C 2023 Skilful forecasts of summer rainfall in the yangtze river basin from november *Adv. Atmos. Sci.* **40** 2082–91

Bett P E, Martin G M, Dunstone N, Scaife A A, Thornton H E and Li C 2021 Seasonal rainfall forecasts for the yangtze river basin in the extreme summer of 2020 *Adv. Atmos. Sci.* **38** 2212–20

Black E, Slingo J and Sperber K R 2003 An observational study of the relationship between excessively strong short rains in coastal East Africa and Indian Ocean SST *Mon. Weather Rev.* **131** 74–94

Boer G J *et al* 2016 The decadal climate prediction project (DCPP) contribution to CMIP6 *Geosci. Model Dev.* **9** 3751–77

Boer G J and Sospedra-Alfonso R 2019 Assessing the skill of the Pacific decadal oscillation (PDO) in a decadal prediction experiment *Clim. Dyn.* **53** 5763–75

Choi K S, Wu C C and Cha E J 2010 Change of tropical cyclone activity by Pacific-Japan teleconnection pattern in the western North Pacific *J. Geophys. Res.* **115** D19114

Di Capua G, Kretschmer M, Runge J, Alessandri A, Donner R V, van den Hurk B, Vellore R, Krishnan R and Coumou D 2019 Long-lead statistical forecasts of the indian summer monsoon rainfall based on causal precursors *Weather Forecast.* **34** 1377–94

Ding Y and Chan J C L 2005 The East Asian summer monsoon: an overview *Meteorol. Atmos. Phys.* **89** 117–42

Ding Y, Dong S I, Liu Y, Wang Z, Li Y, Zhao L and Song Y 2018 On the characteristics, driving forces and inter-decadal variability of the East Asian summer monsoon *Chin. J. Atmos. Sci.* **42** 533–58 (in Chinese)

Ding Y, Liang P, Liu Y and Zhang Y 2020 Multiscale variability of meiyu and its prediction: a new review *J. Geophys. Res. Atmos.* **125** e2019JD031496

Ding Y, Liu Y and Hu Z-Z 2021 The record-breaking Mei-yu in 2020 and associated atmospheric circulation and tropical SST anomalies *Adv. Atmos. Sci.* **38** 1980–93

Ding Y, Wang Z and Sun Y 2008 Inter-decadal variation of the summer precipitation in East China and its association with decreasing Asian summer monsoon. Part I: observed evidences *Int. J. Climatol.* **28** 1139–61

Ester M, Kriegel H-P, Sander J and Xu X 1996 A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining* (*Portland, Oregon*) (*AAAI Press*)

Flato G *et al* 2013 Evaluation of climate models *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* ed T F Stocker, D Qin, G-K Plattner, M Tignor, S K Allen, J Boschung, A Nauels, Y Xia, V Bex and P M Midgley (Cambridge)

He C, Zhou T, Zhang L, Chen X and Zhang W 2023 Extremely hot East Asia and flooding western South Asia in the summer of 2022 tied to reversed flow over Tibetan Plateau *Clim. Dyn.* **61** 2103–19

Hersbach H *et al* 2020 The ERA5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049

Hodges K, Cobb A and Vidale P L 2017 How well are tropical cyclones represented in reanalysis datasets? *J. Clim.* **30** 5243–64

Huang B, Thorne P W, Banzon V F, Boyer T, Chepurin G, Lawrimore J H, Menne M J, Smith T M, Vose R S and Zhang H-M 2017 Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons *J. Clim.* **30** 8179–205

Kim J-S, Li R C-Y and Zhou W 2012 Effects of the Pacific-Japan teleconnection pattern on tropical cyclone activity and extreme precipitation events over the Korean peninsula *J. Geophys. Res. Atmos.* **117** D18109

Kong W and Chiang J C H 2020 Interaction of the westerlies with the tibetan plateau in determining the Mei-Yu termination *J. Clim.* **33** 339–63

Kretschmer M, Runge J and Coumou D 2017 Early prediction of extreme stratospheric polar vortex states based on causal precursors *Geophys. Res. Lett.* **44** 8592–600

Li C, Scaife A A, Lu R, Arribas A, Brookshaw A, Comer R E, Li J, MacLachlan C and Wu P 2016 Skillful seasonal prediction of Yangtze river valley summer rainfall *Environ. Res. Lett.* **11** 094002

Li R C Y, Zhou W and Li T 2013 Influences of the Pacific–Japan teleconnection pattern on synoptic-scale variability in the western North Pacific *J. Clim.* **27** 140–54

Li R K K, Tam C Y, Lau N C, Sohn S J and Ahn J B 2020 Potential predictability of the silk road pattern and the role of SST as inferred from seasonal hindcast experiments of a coupled climate model *J. Clim.* **33** 9567–80

Liu Y and Ding Y 2008 Teleconnection between the Indian summer monsoon onset and the Mei-yu over the Yangtze River Valley *Sci. China* D **51** 1021–35

Lu R Y 2002 Indices of the summertime western North Pacific subtropical high *Adv. Atmos. Sci.* **19** 1004–28

Mantua N J, Hare S R, Zhang Y, Wallace J M and Francis R C 1997 A pacific interdecadal climate oscillation with impacts on salmon production *Bull. Am. Meteorol. Soc.* **78** 1069–80

Martin G M, Dunstone N J, Scaife A A and Bett P E 2020 Predicting june mean rainfall in the middle/lower Yangtze River Basin *Adv. Atmos. Sci.* **37** 29–41

Ng K S-C, Lee M H and Zong Y 2020 A parameter for quantifying the macroscale asymmetry of tropical cyclone cloud clusters *J. Atmos. Oceanic Technol.* **37** 1603–22

Ng K S and Leckebusch G C 2021 A new view on the risk of typhoon occurrence in the western North Pacific *Nat. Hazards Earth Syst. Sci.* **21** 663–82

Ng K S, Leckebusch G C and Hodges K I 2022 A Causality-guided statistical approach for modeling extreme Mei-yu rainfall based on known large-scale modes—A pilot study *Adv. Atmos. Sci.* **39** 1925–40

Ning L, Liu J and Wang B 2017 How does the South Asian High influence extreme precipitation over eastern China? *J. Geophys. Res. Atmos.* **122** 4281–98

Ninomiya K and Shibagaki Y 2007 Multi-scale features of the Mei-yu-baiu front and associated precipitation systems *J. Meteorol. Soc. Jpn.* II **85B** 103–22

Nitta T 1987 Convective activities in the tropical western Pacific and their impact on the Northern Hemisphere summer circulation *J. Meteorol. Soc. Jpn.* II **65** 373–90

NOAA 2022 *The NCEI PDO index* (available at: www.ncei.noaa.gov/pub/data/cmb/ersst/v5/index/ersst.v5.pdo.dat) (Accessed 7 September 2022)

Osinski R, Lorenz P, Kruschke T, Voigt M, Ulbrich U, Leckebusch G C, Faust E, Hofherr T and Majewski D 2016 An approach to build an event set of European windstorms based on ECMWF EPS *Nat. Hazards Earth Syst. Sci.* **16** 255–68

Runge J 2020 Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets Proc. 6th Conf. on Uncertainty in Artificial Intelligence (UAI) ed P Jonas and S David (Proceedings of Machine Learning Research (PMLR)

Runge J, Nowack P, Kretschmer M, Flaxman S and Sejdinovic D 2019 Detecting and quantifying causal associations in large nonlinear time series datasets *Sci. Adv.* **5** eaau4996

Saji N H, Goswami B N, Vinayachandran P N and Yamagata T 1999 A dipole mode in the tropical Indian Ocean *Nature* **401** 360–3

Sampe T and Xie S-P 2010 Large-scale dynamics of the meiyu-baiu rainband: environmental forcing by the westerly jet *J. Clim.* **23** 113–34

Scaife A A *et al* 2022 Long-range predictability of extratropical climate and the length of day *Nat. Geosci.* **15** 789–93

Thompson V, Dunstone N J, Scaife A A, Smith D M, Slingo J M, Brown S and Belcher S E 2017 High risk of unprecedented UK rainfall in the current climate *Nat. Commun.* **8** 107

Trenberth K E 1997 The definition of el niño *Bull. Am. Meteorol. Soc.* **78** 2771–8

United Nations Department of Economic and Social Affairs, Population Division 2022 World population prospects 2022 *United Nations, Department of Economic and Social Affairs* (Population Division)

Wakabayashi S and Kawamura R 2004 Extraction of major teleconnection patterns possibly associated with the anomalous summer climate in Japan *J. Meteorol. Soc. Jpn.* II **82** 1577–88

Wang B and Fan Z 1999 Choice of South Asian summer monsoon Indices *Bull. Am. Meteorol. Soc.* **80** 629–38

Wang B, Liu J, Yang J, Zhou T and Wu Z 2009 Distinct principal modes of early and late summer rainfall anomalies in East Asia *J. Clim.* **22** 3864–75

Wang B, Wu R and Lau K M 2001 Interannual variability of the Asian summer monsoon: contrasts between the Indian and the Western North Pacific–East Asian monsoons *J. Clim.* **14** 4073–90

Wang B, Wu Z, Li J, Liu J, Chang C-P, Ding Y and Wu G 2008 How to measure the strength of the East Asian summer monsoon *J. Clim.* **21** 4449–63

Webster P J and Yang S 1992 Monsoon and ENSO: selectively interactive systems *Q. J. R. Meteorol. Soc.* **118** 877–926

Williams K D *et al* 2018 The Met Office global coupled model 3.0 and 3.1 (GC3.0 and GC3.1) configurations *J. Adv. Model. Earth Syst.* **10** 357–80

Wu J and Gao X-J 2013 A gridded daily observation dataset over China region and comparison with the other datasets *Chin. J. Geophys.* **56** 1102–11

Wu P, Clark R, Furtado K, Xiao C, Wang Q and Sun R 2023 A case study of the July 2021 Henan extreme rainfall event: from weather forecast to climate risks *Weather Clim. Extremes* **40** 100571

Wu R, Hu Z-Z and Kirtman B P 2003 Evolution of ENSO-related rainfall anomalies in East Asia *J. Clim.* **16** 3742–58

Xu Y, Gao X, Shen Y, Xu C, Shi Y and Giorgi F 2009 A daily temperature dataset over China and its application in validating a RCM simulation *Adv. Atmos. Sci.* **26** 763–72

Ye H and Lu R Y 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. Clim.* **24** 2271–84

Zhou B and Wang H 2006 Relationship between the boreal spring Hadley circulation and the summer precipitation in the Yangtze River valley *J. Geophys. Res. Atmos.* **111** D16109

Zhou T, Zhang W, Zhang L, Clark R, Qian C, Zhang Q, Qiu H, Jiang J and Zhang X 2022 2021: a year of unprecedented climate extremes in Eastern Asia, North America, and Europe *Adv. Atmos. Sci.* **39** 1598–607

Zhou Z-Q, Xie S-P and Zhang R 2021 Historic Yangtze flooding of 2020 tied to extreme Indian Ocean conditions *Proc. Natl Acad. Sci.* **118** e2022255118