Ensemble skill gains obtained from the multi-physics versus multi-model approaches for continental-scale hydrological simulations

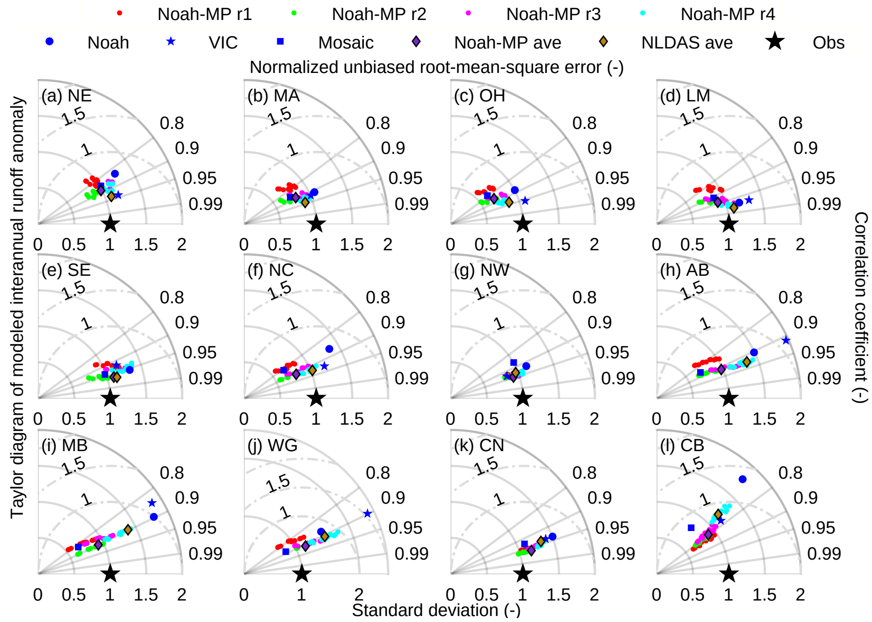
Wenli Fei[1], Hui Zheng[1], Zhongfeng XU[1], Wen-Ying Wu[2], Peirong Lin[3], Ye Tian[4], Mengyao Guo[5], Dunxian She[5], Lingcheng Li[6], Kai Li[1], and Zong-Liang Yang[2]

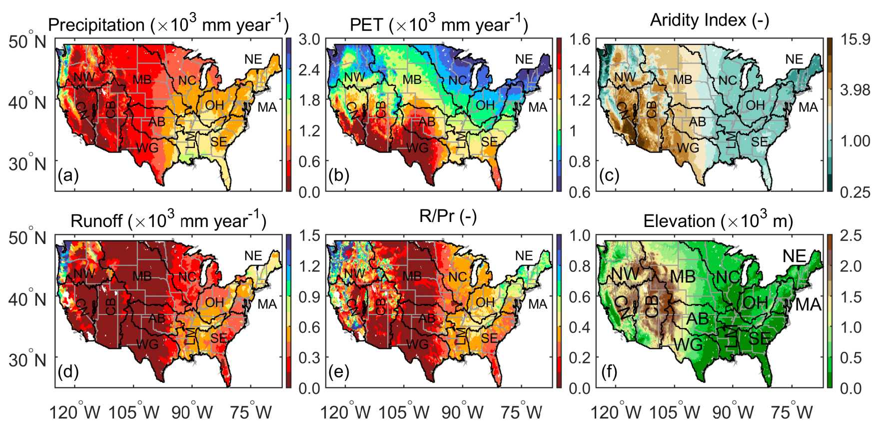
[1]Institute of Atmospheric Physics, Chinese Academy of Sciences
[2]University of Texas at Austin
[3]Princeton University
[4]Nanjing University of Information Science and Technology
[5]Wuhan University
[6]Jackson School of Geosciences, The University of Texas at Austin

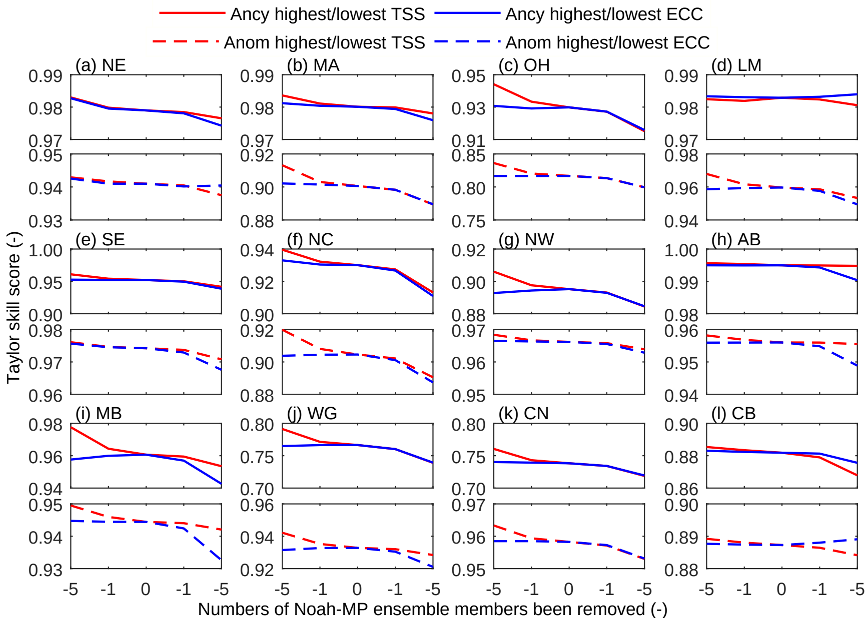
November 22, 2022

Abstract

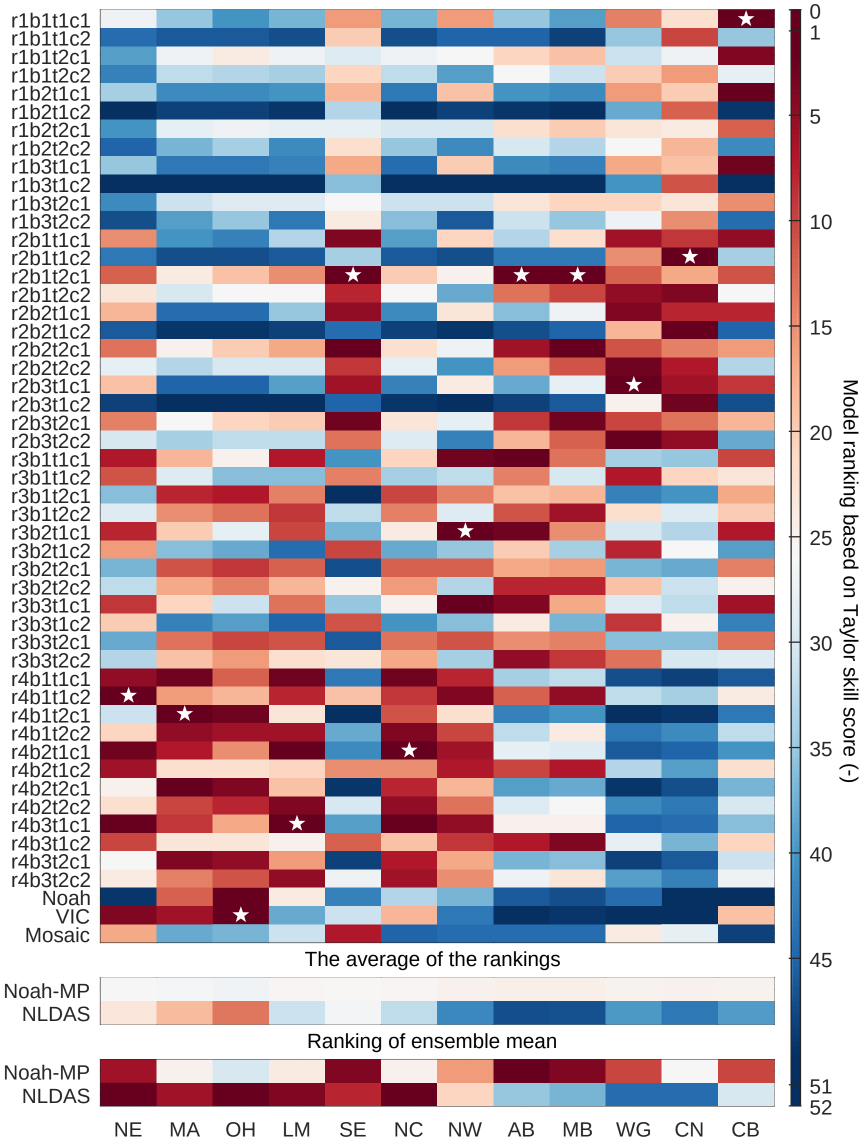
Multi-physics ensemble simulations have emerged as a promising approach to ensemble hydrological simulations due to the advantages in process understanding and model development. As a multi-physics ensemble is constructed by perturbing the physics of multi-physics models, the ensemble members share a substantial portion of the same physics and hence are not independent of each other. It is unknown whether and to what extent the independence of the ensemble members affects the ensemble skill gain, especially compared with the multi-model ensemble approach. This study compares a multi-physics ensemble constructed from the Noah land surface model with multi-parameterization options (Noah-MP) with the North American Land Data Assimilation System (NLDAS) multi-model ensemble. The two ensembles are evaluated at 12 River Forecast Centers over the conterminous United States. The ensemble skill gain is measured by the difference between the performance of the ensemble mean and the average of the ensemble members' performance, and the inter-member independence is measured by error correlations. The results show that the Noah-MP members outperform, on average, the NLDAS models, especially in the snow-dominated areas. In addition, the best-performing models among the two ensembles are mostly Noah-MP members. However, these two performance superiorities do not lead to the superiority of the ensemble mean. The Noah-MP multi-physics ensemble has a low ensemble skill gain, resulting from a high error correlation among the ensemble members. This study suggests that the methods of ensemble construction and optimization should be improved to also consider inter-member independence, especially for a multi-physics ensemble.

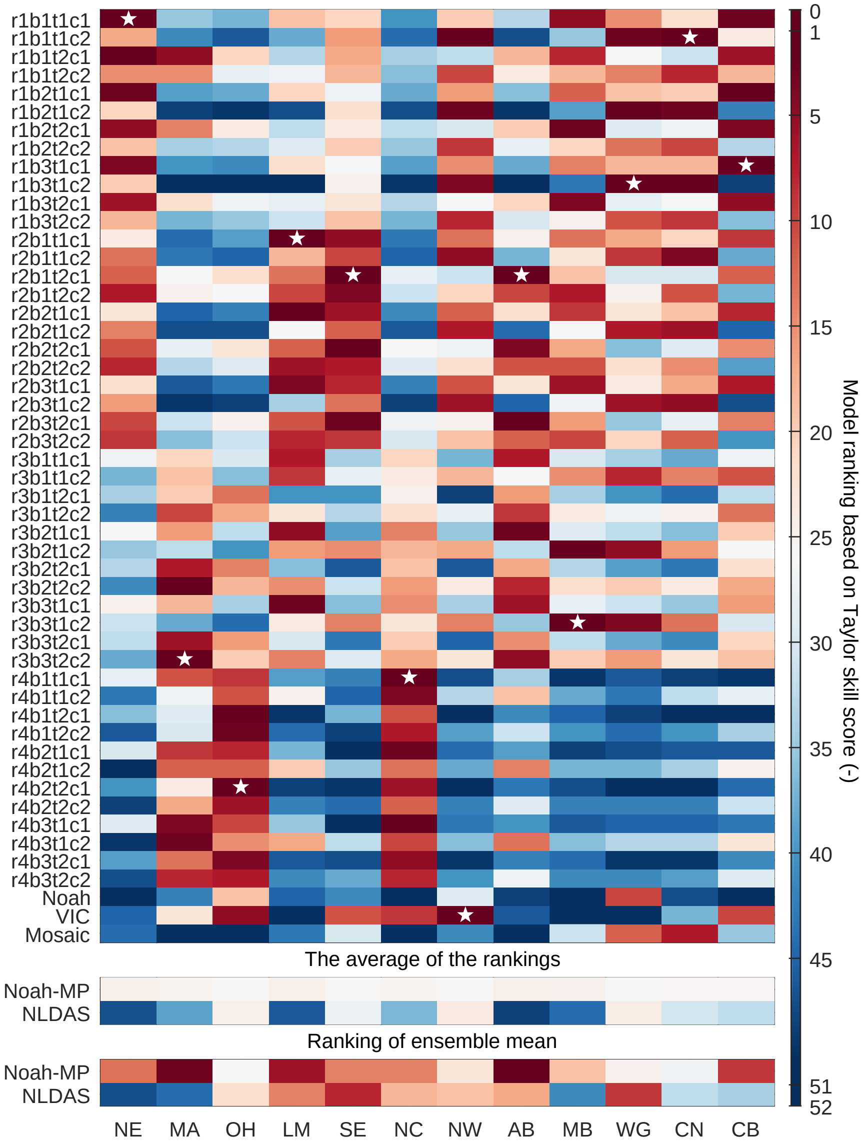

Noah-MP r1
Noah-MP r2
Noah-MP r3
Noah-MP r4
Noah
VIC
Mosaic
Noah-MP ave
NLDAS ave
Obs
Normalized unbiased root-mean-square error (-)
Taylor diagram of modeled interannual runoff anomaly
Correlation coefficient (-)
Standard deviation (-)
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB


Precipitation (×10^3 mm year^{-1})
PET (×10^3 mm year^{-1})
Aridity Index (-)
Runoff (×10^3 mm year^{-1})
R/Pr (-)
Elevation (×10^3 m)
50°N
40°N
30°N
120°W 105°W 90°W 75°W
(a)
(b)
(c)
(d)
(e)
(f)



Ancy highest/lowest TSS
Ancy highest/lowest ECC
Anom highest/lowest TSS
Anom highest/lowest ECC
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB
Taylor skill score (-)
Numbers of Noah-MP ensemble members been removed (-)


r1b1t1c1
r1b1t1c2
r1b1t2c1
r1b1t2c2
r1b2t1c1
r1b2t1c2
r1b2t2c1
r1b2t2c2
r1b3t1c1
r1b3t1c2
r1b3t2c1
r1b3t2c2
r2b1t1c1
r2b1t1c2
r2b1t2c1
r2b1t2c2
r2b2t1c1
r2b2t1c2
r2b2t2c1
r2b2t2c2
r2b3t1c1
r2b3t1c2
r2b3t2c1
r2b3t2c2
r3b1t1c1
r3b1t1c2
r3b1t2c1
r3b1t2c2
r3b2t1c1
r3b2t1c2
r3b2t2c1
r3b2t2c2
r3b3t1c1
r3b3t1c2
r3b3t2c1
r3b3t2c2
r4b1t1c1
r4b1t1c2
r4b1t2c1
r4b1t2c2
r4b2t1c1
r4b2t1c2
r4b2t2c1
r4b2t2c2
r4b3t1c1
r4b3t1c2
r4b3t2c1
r4b3t2c2
Noah
VIC
Mosaic
The average of the rankings
Noah-MP
NLDAS
Ranking of ensemble mean
Noah-MP
NLDAS
NE
MA
OH
LM
SE
NC
NW
AB
MB
WG
CN
CB
0
1
5
10
15
20
25
30
35
40
45
51
52
Model ranking based on Taylor skill score (-)



r1b1t1c1
r1b1t1c2
r1b1t2c1
r1b1t2c2
r1b2t1c1
r1b2t1c2
r1b2t2c1
r1b2t2c2
r1b3t1c1
r1b3t1c2
r1b3t2c1
r1b3t2c2
r2b1t1c1
r2b1t1c2
r2b1t2c1
r2b1t2c2
r2b2t1c1
r2b2t1c2
r2b2t2c1
r2b2t2c2
r2b3t1c1
r2b3t1c2
r2b3t2c1
r2b3t2c2
r3b1t1c1
r3b1t1c2
r3b1t2c1
r3b1t2c2
r3b2t1c1
r3b2t1c2
r3b2t2c1
r3b2t2c2
r3b3t1c1
r3b3t1c2
r3b3t2c1
r3b3t2c2
r4b1t1c1
r4b1t1c2
r4b1t2c1
r4b1t2c2
r4b2t1c1
r4b2t1c2
r4b2t2c1
r4b2t2c2
r4b3t1c1
r4b3t1c2
r4b3t2c1
r4b3t2c2
Noah
VIC
Mosaic
The average of the rankings
Noah-MP
NLDAS
Ranking of ensemble mean
Noah-MP
NLDAS
NE
MA
OH
LM
SE
NC
NW
AB
MB
WG
CN
CB
0
1
5
10
15
20
25
30
35
40
45
51
52
Model ranking based on Taylor skill score (-)

Ensemble skill gains obtained from the multi-physics versus multi-model approaches for continental-scale hydrological simulations

Wenli Fei[1,2], Hui Zheng[1], Zhongfeng Xu[1], Wen-Ying Wu[3], Peirong Lin[4], Ye Tian[5], Mengyao Guo[6], Dunxian She[6], Lingcheng Li[3], Kai Li[1], Zong-Liang Yang[3]

[1]Key Laboratory of Regional Climate-Environment Research for Temperate East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]Department of Geological Sciences, John A. and Katherine G. Jackson School of Geosciences, the University of Texas at Austin, Austin, Texas, USA

[4]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA

[5]School of Hydrology and Water Resources, Nanjing University of Information Science and Technology, Nanjing, China

[6]State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, China

Corresponding authors: Hui Zheng (hzheng_iap@outlook.com) and Zong-Liang Yang (liang@jsg.utexas.edu)

Key Points:

- A significant overlap of the physics among the multi-physics ensemble members leads to a lower inter-member independence and ensemble skill gain

- The performance of the ensemble mean responds asymmetrically to the inclusion of an independent versus a non-independent member
- An ensemble averaging and optimization method that can account for the inter-member independence is needed to maximize the multi-physics ensemble skill gain

Abstract

Multi-physics ensemble simulations have emerged as a promising approach to ensemble hydrological simulations due to the advantages in process understanding and model development. As a multi-physics ensemble is constructed by perturbing the physics of multi-physics models, the ensemble members share a substantial portion of the same physics and hence are not independent of each other. It is unknown whether and to what extent the independence of the ensemble members affects the ensemble skill gain, especially compared with the multi-model ensemble approach. This study compares a multi-physics ensemble constructed from the Noah land surface model with multi-parameterization options (Noah-MP) with the North American Land Data Assimilation System (NLDAS) multi-model ensemble. The two ensembles are evaluated at 12 River Forecast Centers over the conterminous United States. The ensemble skill gain is measured by the difference between the performance of the ensemble mean and the average of the ensemble members' performance, and the inter-member independence is measured by error correlations. The results show that the Noah-MP members outperform, on average, the NLDAS models, especially in the snow-dominated areas. In addition, the best-performing models among the two ensembles are mostly Noah-MP members. However, these two performance superiorities do not lead to the superiority of the ensemble mean. The Noah-MP multi-physics ensemble has a low ensemble skill gain, resulting from a high error correlation among the ensemble members. This study suggests that the methods of ensemble construction and optimization should be improved to also consider inter-member independence, especially for a multi-physics ensemble.

1 Introduction

Multi-model ensemble simulations have been broadly shown to offer a systematic improvement over individual models (Georgakakos et al., 2004; Shamseldin et al., 1997). Models are remarkably different from each other in parameterizing various hydrological processes. No single model exhibits clear superiority in all terrestrial water fluxes under all climatic conditions. By combining multiple models, Gao and Dirmeyer (2006), Gudmundsson et al. (2012b), Xia et al. (2012a), and Beck et al. (2017) showed that the arithmetic average of the ensemble outperforms all or most of the constituent members. This superiority is reflected as an asymmetric response of the ensemble mean to the inclusion of better- versus worse-performing models: there is an improvement when a better model is included, but little or no apparent degradation when a worse model is added (Ajami et al., 2006; Guo et al., 2007). This asymmetry indicates that the performance of the ensemble mean (PEM) differs from the arithmetic average of the ensemble members' performance (AEP).

"The key to the success of the multi-model concept lies in combining independent and skillful models, each with its own strengths and weaknesses" (Hagedorn et al., 2005). With independent models, the errors associated with individual models can somehow cancel each other out, leading to a superior-performing ensemble mean. This importance of inter-model independence has been clearly shown in various analyses. Yoo and Kang (2005) measured the performance by the error correlation coefficient. Using mathematical decomposition of the correlation coefficient, they showed that the difference between the PEM and the AEP increases if the errors of the ensemble members are less correlated with each other. Winter and Nychka (2009) represented model errors as vectors in T dimensions (where T is the number of time steps), the length of which is the root mean square error. Their geometric analyses clearly showed that the PEM reaches a maximum (i.e., mean square error reaches a minimum) if the set of member models can sample the error space evenly. From the perspective of information theory (Goodwell et al., 2020; Goodwell & Kumar, 2017a, 2017b), non-independent models contain redundant information. On the basis of a measure of mutual information content, Sharma et al. (2019) demonstrated that the skill gains obtained by multi-model ensembles are dominated by model independence.

The North American Land Data Assimilation System (NLDAS) (Mitchell et al., 2004) adopts the concept of the multi-model ensemble and runs four distinct models over the conterminous United States (CONUS). The four models in NLDAS phase 2 are the Noah land surface model (Noah, version 2.8), Mosaic, the Variable Infiltration Capacity (VIC, version 4.0.3) model, and the Sacramento Soil Moisture Accounting Model (SAC). These NLDAS models differ remarkably in model structure and process parameterization (Kumar et al., 2017; Xia et al., 2012b). Using confirmatory factor analysis, Kumar et al. (2017) confirmed that these structural differences result in dissimilar model behaviors, especially in the simulations of runoff. However, with a long history of extensive evaluations and improvements (Xia et al., 2019; Xia et al., 2012a; Xia et al., 2012b), the performances of these NLDAS models are shown to be high. They can well replicate the observed evapotranspiration (ET) (Long et al., 2014; Zhang et al., 2020), streamflow (Xia et al., 2012a), soil moisture (Xia et al., 2015), and groundwater (Xia et al., 2017) patterns over the CONUS. The sound inter-model independence and satisfactory model performance make the NLDAS multi-model ensemble an ideal benchmark for new ensemble techniques at a continental scale.

Recently, multi-physics models have emerged as a new tool for performing ensemble simulations. The multi-physics models host different parameterization schemes for several key processes (M. P. Clark et al., 2011). An ensemble of model configurations can be generated by selecting different combinations of parameterization schemes given land surface processes (Gan et al., 2019; Yang et al., 2011; Zhang et al., 2016; Zheng & Yang, 2016). Numerous studies using this kind of multi-physics ensemble simulation have already been conducted (Clark et al., 2010; Coxon et al., 2014; Krueger et al., 2010; McMillan et al., 2010; Oudin et al., 2006), addressing various research questions. By discriminating competing model parameterizations, the linkage between model parameterizations and catchment type and hydrological signatures can be established (Clark et al., 2010; Clark et al., 2016; McMillan et al., 2010). Such analyses improve our understanding of the dominant hydrological processes in various catchments and hydrological events (Coxon et al., 2014; Douinot et al., 2018). The multi-physics ensemble can also benefit uncertainty attribution and model development. The impacts of a specified process on the overall model behavior can be pinpointed, and the interplays between different processes can be disentangled (Li et al., 2019; You et al., 2020; Zhang et al., 2016; Zheng et al., 2019). Recognizing these advantages in process understandings and uncertainty attribution, several mainstream models have adopted this concept. Available multi-physics models include the Noah land surface model with multi-parameterization options (Noah-MP) (Niu et al., 2011), the Community Land Model version 5 (CLM5) (Lawrence et al., 2019), the Structure for Unifying Multiple Modeling Alternatives (SUMMA) (Clark et al., 2015a; Clark et al., 2015b), and the Joint UK Land Environment Simulator (JULES) (Best et al., 2011; D. B. Clark et al., 2011).

Among these available multi-physics models, Noah-MP has been broadly tested at a variety of spatiotemporal scales. Noah-MP (Niu et al., 2011; Yang et al., 2011) is augmented over the Noah model by improving the physical realism of the parameterization of snow (Yang et al., 2011), vegetation canopy (Dickinson et al., 1998), groundwater (Niu et al., 2007), and frozen soil (Niu & Yang, 2006). Various configurations of the Noah-MP model have been broadly evaluated. Yang et al. (2011) and Cai et al. (2014a) verified that the physical realism improves the model performance in simulating runoff, groundwater, soil moisture, snow, and total water storage (TWS) at the basin and global scales. Ma et al. (2017) reported that Noah-MP is more skillful than the NLDAS models in simulating runoff. Xia et al. (2016a) and Cai et al. (2014b) also showed that Noah-MP outperforms NLDAS models in capturing the seasonal cycle of snow and snowmelt runoff due to the inclusion of a multilayer snowpack and the improvement of the turbulence parameterization. The credibility obtained from the improved physical realism and model performance promotes the adoption of Noah-MP in large-scale operational systems, including the Weather Research and Forecasting (WRF) model (Barlage et al., 2015) for weather forecasts and the National Water Model (NWM) (Maidment, 2017) for hydrological forecasts. The model is also undergoing extensive tests for the next phase of NLDAS (Xia et al., 2017; Zhang et al., 2020). Any improvement in the simulation performance from Noah-MP would directly benefit these operational systems.

However, it is largely unknown how the emerging multi-physics approach compares with the well-established multi-model approach for ensemble hydrological simulations. It is often hypothesized that the multi-physics approach has a performance advantage. Multi-physics ensembles provide a broad coverage of the feasible model physics space (M. P. Clark et al., 2011). There is likely a multi-physics ensemble member that can best approximate the reality. The inclusion of such a superior-performing member can result in a superior PEM, as a consequence of its asymmetric response to the constituent's performance. Furthermore, the best-performing multi-physics member varies with basins and climatic conditions (Gan et al., 2019). A multi-physics ensemble that includes these members can offer a systematic performance improvement under various conditions. However, these advantages may be offset by a hidden pitfall. The multi-physics ensemble members share a substantial portion of the same parameterization schemes with each other. These physics overlaps reduce the inter-member independence and may lower the ensemble skill gains. This negative effect has been greatly outlooked. A comprehensive comparison of the multi-physics and multi-model ensemble approaches is needed.

This study presents such a comparison between the Noah-MP multi-physics ensemble and the NLDAS multi-model ensemble. The comparisons are performed for the 12 River Forecast Centers (RFCs) over the CONUS. This area covers a wide variety of climatic regimes, which could ensure the robustness of the comparison (Gupta et al., 2014). In the remainder of this paper, section 2 details the two ensembles and the observations. Section 3 describes the definition of ensemble skill gain and skill metrics. Section 4 shows the results with discussion. Section 5 provides conclusions.

2 Model and data

2.1 The NLDAS multi-model ensemble

We use three NLDAS models: Noah-2.8, VIC-4.0.3, and Mosaic. They are the only three NLDAS models whose outputs are publicly available from NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) (https://disc.gsfc.nasa.gov/datasets?keywords=NLDAS). A brief introduction of these three models is as follows.

Noah (Ek et al., 2003) is developed as the land component of the weather and climate forecasting models of Nation Centers for Environmental Prediction (NCEP) and National

Oceanic and Atmospheric Administration (NOAA). Noah has four soil layers and one snow layer. There is a dominant vegetation type for each grid cell. In terms of model physics, Noah considers a comprehensive time-dependent canopy resistance (Chen et al., 1996), seasonal frozen soils, and the snow accumulation/ablation processes (Koren et al., 1999). Noah 2.8 (the version used in NLDAS-2 and this study) has made many improvements based on Noah 2.7.1 (the version used in NLDAS-1), including modification of the albedo formulation by combining snow-albedo decay and liquid-water refreeze (Livneh et al., 2010) and the introduction of seasonal factors to the simulation of warm-season processes (Wei et al., 2013).

Mosaic (Koster & Suarez, 1992) is a land surface model developed for use within the general circulation model (GCM) (Ducharne et al., 1999). It includes three soil layers and one snow layer. Mosaic features the “mosaic” strategy for considering the surface heterogeneity. There are several vegetation tiles within each model grid cell, where energy and water balance are calculated separately (Yang et al., 2003). The calculations are based on the Simple Biosphere (SiB) model (Sellers et al., 1986), which is analogous to the electrical resistance method in calculating the energy and matter transfer in biophysical processes.

VIC (Liang et al., 1994) is a widely used semi-distributed hydrology model with a full Surface Vegetation-Atmosphere Transfer (SVAT) representation. It features a variable infiltration capacity approach for parameterizing runoff generation. VIC includes three soil layers and two snow layers. Like Mosaic, VIC also considers several vegetation types with a tiling method (Cherkauer et al., 2003). Furthermore, VIC utilizes subgrid elevation bands to realistically describe the dependence of temperature, precipitation, and snow on elevation in the snow-dominated regions (Liang et al., 1994). The version used in this study, 4.0.3, is subject to the constraints of both surface water and energy conservations.

All the NLDAS models are driven by a set of high-quality atmospheric forcings, topography, vegetation, and soil types. The spatial resolution of these datasets is 0.125° × 0.125°. The NLDAS-2 total precipitation field combines the gauge-based precipitation from NOAA Climate Prediction Center (CPC) with the monthly Parameter-elevation Regressions on Independent Slopes Model (PRISM) topographical adjustment, Doppler Stage II radar precipitation data, NOAA CPC Morphing Technique data (CMORCH), and the NARR precipitation. The non-precipitation forcings are derived from the NCEP North American Regional Reanalysis (NARR) analysis fields (Cosgrove et al., 2003). The surface downward shortwave radiation data are also bias-corrected by the hourly 1/8th-degree GOES-based surface downward shortwave radiation fields (1996-2000) (Pinker et al., 2003). The NLDAS topography is based on the GTOPO30 Global 30 Arc Second (~1 km) Elevation Dataset. The 14-class UMD map at 1 km of the University of Maryland's UMD Land Cover Classification is used as the NLDAS vegetation class dataset. The monthly leaf area index (LAI) dataset is re-gridded from NOAA National Environmental Satellite, Data, and Information Service (NESDIS) 0.144° monthly climatology LAI (Gutman & Ignatov, 1998). The NLDAS soil data with 16 texture types over the CONUS is derived from 1 km Penn State STATSGO data. These datasets have been widely used and tested (Cai et al., 2014a; Cai et al., 2014b; Xia et al., 2016a; Xia et al., 2016b; Xia et al., 2015).

2.2 The Noah-MP multi-physics ensemble

Forty-eight configurations of Noah-MP version 3.6 are constructed by perturbing the model physics of runoff generation, stomatal conductance, soil moisture limitation factor to transpiration (β-factor), and turbulence. These processes are selected based on their importance suggested in previous studies (Yang et al., 2011; Zhang et al., 2016; Zheng et al., 2019). There are two distinct stomatal conductance schemes (Ball–Berry, Jarvis), three β-factor schemes (NOAHB, CLM, SSiB), four runoff schemes (SIMGM, SIMTOP, NOAHR, BATS), and two turbulence schemes (M-O, Chen97) (48 = 2 × 3 × 4 × 2). The four runoff schemes fall into two groups. The first group consists of SIMGM and SIMTOP. They are TOPMODEL-based schemes with a groundwater component. The groundwater in SIMGM is dynamic, which interacts with the soil moisture, whereas the groundwater in SIMTOP is determined by the bottom soil moisture based on an equilibrium assumption. The second group, namely NOAHR and BATS, does not have a groundwater component. NOAHR represents the infiltration excess runoff-generation mechanism, whereas BATS represents the idea of fractional saturation area. Details of the parameterization schemes can be found in Table 1 of Zheng et al. (2019).

The Noah-MP multi-physics ensemble is driven by the same atmospheric forcings and static inputs as the NLDAS models. The simulations span 30 years from January 1982 to December 2011. The initial condition on 1 January 1980 is generated by looping the 1979 simulations 100 times.

2.3 The USGS HUC8 runoff data

The monthly 8-digit Hydrologic Unit Code (HUC8) runoff data from 1982 to 2011 are used as the observations. The data are derived from daily stream-gauge observations by the USGS. The derivation is based on two assumptions: (1) the runoff is uniform within each HUC8; (2) the river routing can be ignored because the propagation of flow waves at basin scale is in days (Allen et al., 2018). With these two assumptions, the HUC8 runoff data are weighted across the stream-gauge observations within each HUC8, where the weight is the overlap area of the gauge-observed basins and the HUC8 basin. The USGS HUC8 runoff data have been widely selected to evaluate runoff simulation at regional scales (Ma et al., 2017; Zheng et al., 2019).

2.4 RFCs over the CONUS and their climatology

The above-mentioned simulations and observations are upscaled into 12 RFCs over the CONUS, namely Northeast (NE), Mid-Atlantic (MA), Ohio (OH), Lower Mississippi (LM), Southeast (SE), North Central (NC), Northwest (NW), Arkansas (AB), Missouri (MB), West Gulf (WG), California-Nevada (CN), and Colorado (CB). As discussed in Gudmundsson et al. (2012a) and Gudmundsson et al. (2012b), the spatial aggregation of the observations/simulations from smaller basins (HUC8 and NLDAS grid cells) reduces the measurement errors in observations and modeling errors in spatially varying parameters. Furthermore, as the observations are taken in relatively smaller basins (i.e., HUC8) and used for a relatively long timescale (i.e., monthly), river routing is not necessary, which may also introduce modeling errors.

Figure 1 shows the boundaries and climatology, including multi-year averaged precipitation, potential evapotranspiration, runoff, runoff ratio, and Budyko’s aridity index. The potential evapotranspiration is calculated from the NLDAS daily forcings based on the FAO-56 algorithm (Allen et al., 1998) of the Penman–Monteith equation (Monteith, 1965; Penman, 1956). Budyko’s aridity index (Budyko, 1974) is defined as the ratio of potential evapotranspiration over precipitation. The RFC-averaged values of the climatology are shown in Table S1. Consistent with Budyko’s hypothesis, the runoff ratio generally decreases with increasing aridity. However, three abnormalities exist in NW, CN, and CB. These abnormally high values hint at possible precipitation underestimation and/or runoff overestimation.

2.5 Annual cycle and interannual anomaly

The annual cycle and interannual anomaly are evaluated separately. The decomposition of the modeled and observed runoff into multi-year averaged climatology (r_{clim}), mean annual cycle ($r_{ancy,m}$), and interannual anomaly ($r_{anom,y,m}$) is as follows. For a modeled or observed runoff $r_{y,m}$ for the *m*th month ($m = 1 \ldots 12$) of the *y*th year ($y = 1 \ldots Y$), where $Y = T/12$

$$r_{clim} = \frac{1}{12Y}\sum_{y=1}^{Y}\sum_{m=1}^{12} r_{y,m} \tag{1}$$

$$r_{ancy,m} = \frac{1}{Y}\sum_{y=1}^{Y} r_{y,m} - r_{clim} \tag{2}$$

$$r_{anom,y,m} = r_{y,m} - r_{ancy,m} - r_{clim} \tag{3}$$

The mean annual cycle gives insights into seasonal variation. The interannual anomaly reflects year-to-year variation, which reflects a model's responses to the monthly perturbations in atmospheric forcings.

3 Ensemble skill gain and skill measures

3.1 Ensemble skill gain

For an ensemble of N members ($x^{(i)}, i = 1 \cdots N$), we define the ensemble skill gain (G) as the difference between the PEM ($S(\bar{x})$) and the AEP ($\overline{S(x)}$).

$$G = S(\bar{x}) - \overline{S(x)} \tag{4}$$

$$\bar{x} = \sum_{i=1}^{N} w_i x^{(i)} \tag{5}$$

$$\overline{S(x)} = \sum_{i=1}^{N} w_i S(x_i) \tag{6}$$

where S denotes a skill measure, which is detailed in the next section. w_i is the weight for $x^{(i)}$ with a constraint of $\sum_{i=1}^{N} w_i = 1$. A larger value of G indicates a large ensemble skill gain.

As broadly reported, the ensemble weights (w_i) have a significant influence on the PEM ($S(\bar{x})$) (Ajami et al., 2006; Bohn et al., 2010), and thus on the ensemble skill gain (G). There are a number of sophisticated weighting methods that can optimize the PEM (Ajami et al., 2007; Arsenault et al., 2015; Duan et al., 2007; Hsu et al., 2009; Marshall et al., 2006; Oudin et al., 2006; Vrugt et al., 2006; Vrugt & Robinson, 2007). However, most of these methods assume inter-model independence, which is not appropriate for this study. Furthermore, the derivation of an optimal set of weights heavily depends on the objective functions, the referencing model signatures, the variables of interests, research catchments, study periods, and reference datasets. These will unnecessarily complicate this study without improving the universality of the conclusions. As results, in this study, we use equal weights (i.e., $w_i = 1/N$), which are also widely used without utilizing prior information (Gao & Dirmeyer, 2006) and shown to be robust under non-stationary conditions (Ajami et al., 2006; Beck et al., 2017; Georgakakos et al., 2004; Guo et al., 2007).

3.2 Skill measures

We quantify the ensemble skill gains based on several different skill measures. However, for conciseness, we summarize the performance mainly based on the Taylor diagram and Taylor skill score (TSS) (Taylor, 2001). The analyses based on Nash–Sutcliffe efficiency (NSE), Kling–

Gupta efficiency (KGE) (Gupta et al., 2009), and correlation coefficient (R) can be found in the supporting information (Tables S2–S5).

The Taylor diagram (Taylor, 2001) is a two-dimensional plot that visually summarizes multiple aspects of the performance of a model simulation (f) relative to the observations (o), including the correlation coefficient (R), normalized unbiased root-mean-square error (nuRMSE), and normalized variability ($\widehat{\sigma_f}$):

$$R = \frac{\frac{1}{T}\sum_{t=1}^{T}(f_t - \bar{o})(r - \bar{o})}{\sigma_f \sigma_o} \tag{7}$$

$$\mathrm{nuRMSE} = \frac{\mathrm{uRMSE}}{\sigma_o} = \frac{1}{\sigma_o}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\left(f_t - \bar{f}\right) - \left(o_t - \bar{o}\right)\right]^2} \tag{8}$$

$$\widehat{\sigma_f} = \frac{\sigma_f}{\sigma_o} = \frac{1}{\sigma_o}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\left(f_t - \bar{f}\right)\right)^2} \tag{9}$$

$$\sigma_o = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\left(o_t - o\right)\right)^2} \tag{10}$$

where T is the total time step number, $\bar{f} = \sum_{t=1}^{T} f_t$ is the temporal mean of the model simulation (f_t), and $\bar{r} = \sum_{t=1}^{T} r_t$ is the temporal average of the observation. σ_f and σ_o are the standard deviation of the model simulation and the observation, respectively. In the Taylor diagram, $\widehat{\sigma_f}$ measures the magnitude of the variation of the model simulation and R measures the timing of the variation.

According to the diagram, the TSS (Taylor, 2001) is defined as follows:

$$\mathrm{TSS} = \frac{4(1+R)}{\left(\frac{\sigma_f}{\sigma_o} + \frac{\sigma_o}{\sigma_f}\right)^2 (1+R_0)} \tag{11}$$

where R_0 is the maximum correlation coefficient attainable. R_0 is assumed to be 1 in this study. TSS ranges from 0 to 1. A higher value indicates a higher consistency between the model prediction and the observations.

3.3 Error correlation

The independence between each two ensemble members is measured by the error correlation coefficient (ECC) as follows.

$$\mathrm{ECC}_{i,j} = \frac{cov(e_i, e_j)}{\sqrt{\sigma_{e_i}}\sqrt{\sigma_{e_j}}} = \frac{\sum_{t=1}^{T}(e_{i,t} - \bar{e}_i)(e_{j,t} - \bar{e}_j)}{\sqrt{\sigma_{e_i}}\sqrt{\sigma_{e_j}}} \tag{12}$$

$$e_t = f_t - o_t \tag{13}$$

where error (e_t) is defined as the difference between the model simulation and the observations; $e_{i,t}$ and $e_{j,t}$ $(i, j = 1 \dots N, i \neq j)$ are the errors of two ensemble members; $\bar{e}_i$ and $\bar{e}_j$ are their temporal means; σ_{e_i} and σ_{e_j} are their standard deviations.

The error correlation coefficient ranges from −1 to +1. A lower error correlation indicates strong independence and can potentially generate a higher ensemble skill gain. If the error correlation between two ensemble members equals −1, their errors can somehow cancel each other out at each time step. The ensemble mean should show superior performance. If two ensemble members have an error correlation of +1, we do not expect an ensemble skill gain from the average of the two ensemble members. As the ensemble size increases beyond two, an ensemble of independent members is expected to have an average error correlation coefficient of 0.

4 Results and Discussion

We individually examine the performance of the ensemble members in section 4.1. Then, in section 4.2, they are ranked for inter-comparisons. The best member and the ensemble mean are identified. Lastly, inter-member independence and its impacts on the PEM are presented and discussed.

4.1. Performance difference within and between the two ensembles

Figure 2 compares the simulated annual runoff cycle from the NLDAS multi-model ensemble and the Noah-MP multi-physics ensemble. The NLDAS models produce different timing of the runoff peaks. The differences are the most notable in NE, NC, NW, and CB. These RFCs are either in the northern CONUS or in mountainous areas, where the spring runoff peak is dominated by snowmelt. Among the NLDAS models, VIC performs the best in capturing the timing of the runoff peak, with a relatively small bias of approximately one month earlier. Such outperformance may be attributable to the detailed consideration of elevation bands (Liang et al., 1994). Noah performs the worst, with a one- to two-month lag in NE and NC and a two-month lead in CB. Compared with the NLDAS models, the Noah-MP ensemble members perform better, especially in the snow-dominated RFCs mentioned above (i.e., NE, NC, NW, and CB). This superiority has been previously reported and considered to be due to the multilayer snowpack physics (Cai et al., 2014b; Ma et al., 2017; Xia et al., 2016a).

Figure 2 also shows the simulated magnitude (or variability) of the annual cycle. The NLDAS models differ from each other remarkably. The Noah model tends to overestimate the variability, whereas the Mosaic model tends to underestimate it. The VIC model lies between these two models and is closest to the observations. Compared with the NLDAS models, the Noah-MP multi-physics ensemble members again show an overall better performance. This outperformance can also be found in other performance criteria, including NSE, KGE, and *R* (Tables S2–S4). It is of interest to notice that the inter-member difference tends to be related to the climatic aridity, especially for the Noah-MP multi-physics ensemble (also shown in Figures 3, 4, and S7). In NE and MA, where the climate is humid, the Noah-MP ensemble members perform similarly to each other. The significant inter-member difference appears in AB, MB, WG, CN, and CB (Figure S7). All these RFCs are spatially adjacent in the southwest CONUS, with a semi-arid or arid climate. The remarkable inter-member differences in arid RFCs may be related to the difficulties in simulating terrestrial water storage anomalies (Cai et al., 2014b).

The performance of the NLDAS models and the Noah-MP ensemble is shown in Figure 3. The NLDAS models scatter significantly, in which the climatic aridity and snow play different roles. The spread in semi-arid and arid RFCs (e.g., AB, MB, WG, CN) are mainly manifested in modeled variability (i.e., spread in the radial direction). The spread in snow-dominated RFCs (e.g., NE, NC, NW) is mainly manifested in the correlation coefficient, which corresponds to the above-mentioned bias in the simulated timing of the spring runoff peak. In CB, where the climate is the driest and where snow is important, the spread is the most pronounced, suggesting a double difficulty in this arid and snow-dominated RFC.

The performance spread among the Noah-MP ensemble members is generally comparable to that among the NLDAS models. In humid RFCs (e.g., NE, MA, OH, LM), the Noah-MP ensemble members are similar to each other and comparable to the observations. The spread increases in arid RFCs and the spread increase is mainly manifested in the modeled variability (i.e., spread in the radial direction). However, there are two notable differences between the Noah-MP and the NLDAS ensembles. First, in snow-dominated RFCs (e.g., NE, NC, NW), the spread in the correlation coefficient among the Noah-MP ensemble members is small. All Noah-MP configurations correlate well with the observations, suggesting the superiority of the multilayer snowpack physics. Second, the Noah-MP ensemble members cluster according to the runoff parameterizations. SIMTOP (r2) tends to have a relatively low variability, whereas BATS (r4) tends to obtain a relatively high variability. SIMGM (r1) separates from the others with a high correlation coefficient in NE and CB but a low correlation coefficient in OH, LM, SE, and AB. It is interesting to note that the SIMTOP scheme (r2) with equilibrium groundwater performs closer to the two runoff parameterization schemes without groundwater (i.e., NOAHR (r3) and BATS (r4)).

Figure 4 shows the Taylor diagram for the interannual anomaly. Compared with the annual cycle (Figure 3), the performance spread is slightly reduced for the Noah-MP ensemble but significantly reduced for the NLDAS models. The decrease in the spread is mainly attributable to the decreases in the correlation coefficient spread, which is the most obvious in snow-dominated RFCs (e.g., NE, NC, NW, CB). This smaller spread in the correlation coefficient suggests a tighter control of the atmospheric forcing on the interannual anomalies than on the annual cycle. The cluster of the runoff parameterization schemes is clearly apparent. The separation is mainly manifested in the correlation coefficient, and SIMGM (r1) is distinguishable from the others. In most RFCs except CN and CB, SIMGM (r1) obtains a distinguishable low correlation coefficient. The other three runoff parameterizations of Noah-MP (i.e., SIMTOP (r2), NOAHR (r3), BATS (r4)) and the three NLDAS models obtain a similar correlation coefficient. They mainly differ in the modeled variability. The BATS runoff scheme (r4) tends to have the largest variability, whereas SIMTOP (r2) tends to have the lowest.

Table 1 summarizes the average TSS obtained from the NLDAS and the Noah-MP ensembles. Their performance deteriorates from the humid to arid RFCs. On average, Noah-MP outperforms NLDAS for both the annual cycle and interannual anomaly. For the annual cycle, the Noah-MP ensemble outperforms the NLDAS ensemble in humid RFCs, but the outperformance is marginal in arid RFCs. In snow-dominated RFCs (e.g., NE, NC, and CB), the NLDAS ensemble further deteriorates, and Noah-MP shows a clearer superiority. For the interannual anomaly, the performance of the Noah-MP ensemble is consistently high in both humid and arid RFCs. The NLDAS ensemble is only marginally worse than Noah-MP in humid RFCs, but the performance deteriorates quickly in arid RFCs. Furthermore, for the interannual variability, snow does not have an obvious impact on the NLDAS models.

4.2. The best member and the ensemble mean

In the previous section, we explored how climatic aridity, snow, and groundwater influence the ensemble performance. Here, we analyze how Noah-MP members and NLDAS models perform across different RFCs. The analyses are based on rankings — by ranking the 48 Noah-MP and the 3 NLDAS ensemble members together, we can identify the best ensemble member.

Figure 5 shows the model rankings for the annual cycle. No single Noah-MP configuration/NLDAS model shows clear superiority in all RFCs. The BATS (r4) runoff option outperforms the other runoff options in MA, OH, and NC. However, it performs the worst in NE, SE, NW, MB, WG, CN, and CB. As shown in Figure 3, the reasons for the underperformance also vary: low correlation and high variability in NE, MB, CN, and CB, and high variability in SE, NW, and WG. SIMGM (r1) outperforms the other runoff options in NE, NW, MB, WG, CN, and CB, but it performs the worst in OH, NC, and AB. Consistent with previous work (Li et al., 2019; Zheng et al., 2019), different parameterizations interplay with each other. The Ball–Berry-type scheme (c1) of the stomatal conductance parameterization outperforms the Jarvis-type scheme (c2) in combination with the SIMGM runoff scheme (r1) at most RFCs except SE, NW, WG, and CN, whereas the Jarvis scheme (c2) outperforms the Ball–Berry scheme (c1) in combination with the NOAHR runoff scheme (r3) in SE, NW, MB, WG, CN, and CB. We also find that the two runoff options with groundwater (i.e., SIMGM (r1) and SIMTOP (r2)) show some superiority in snow-dominated RFCs, including NE, NW, CN, and CB. These interplays among different processes and climatic aridity are not completely understood.

Figure 6 shows the model rankings for the interannual anomaly. Compared with the annual cycle (Figure 5), the SIMGM runoff option (r1) clearly degrades, whereas SIMTOP (r2) and BATS (r4) improve. Note that in SIMGM (r1), runoff is parameterized with the groundwater level, which interacts with soil moisture. In the other schemes, runoff is directly controlled by soil moisture. The improvements of SIMTOP (r2) and BATS (r4) may suggest that soil moisture plays a more important role in mediating the interannual atmospheric anomaly and runoff anomaly. This hypothesis is further supported by the fact that BATS (r4) performs better in humid RFCs, whereas SIMTOP (r2) performs better in arid RFCs. This difference reflects the dependence of the active soil layer depth on climatic aridity: BATS (r4) parameterizes runoff using surface soil moisture, whereas SIMTOP (r2) parameterizes runoff using bottom soil moisture. However, despite the overall underperformance of SIMGM (r1), it still shows superiority in CB. There is still no single Noah-MP configuration or NLDAS model that can outperform all the others in all RFCs.

We also identified the best member among all the members of the two ensembles. Consistent with the hypothesis described in the Introduction, the Noah-MP multi-physics ensemble contains the best member in almost all the RFCs. The VIC model is only marginally better than the best Noah-MP ensemble member in NW for the annual cycle and in OH for the interannual anomaly. As shown in Table 2, the performance of the best ensemble member is remarkably high, with a TSS value higher than 0.9 in all RFCs for both the annual cycle and interannual anomaly. The best ensemble member varies significantly with the RFCs, and no parameterization schemes clearly stand out.

The Noah-MP ensemble has better average performance (Table 1) and always contains the best ensemble members (Table 2). Thus, the Noah-MP ensemble mean should clearly outperform the NLDAS ensemble mean in all RFCs. However, the following results contradict this hypothesis.

Figure 5 also shows the ranking of the ensemble mean for the annual cycle at the bottom. The NLDAS ensemble mean shows comparable or better performance than the Noah-MP ensemble mean in semi-humid to semi-arid RFCs, including OH, LM, SE, NC, NW, AB, and WG. The superiority of the NLDAS ensemble mean is more apparent for the interannual anomaly (Figure 6) than for the annual cycle (Figure 5), where the NLDAS ensemble mean clearly outperforms the Noah-MP ensemble mean in humid to semi-humid RFCs, including NE, MA, OH, LM, SE, and NC. The relatively higher performance of the NLDAS ensemble mean suggests a higher ensemble skill gain obtained from the NLDAS models, which is confirmed in Table 3. The ensemble skill gain measured in NSE and KGE is also shown in Table S5. As discussed in the Introduction, the relatively lower ensemble skill gain obtained from the Noah-MP multi-physics ensemble hints that the multi-physics ensemble members may be too similar to each other.

4.3. Inter-member independence measured by error correlation

Figure 7 shows the error correlation for the Noah-MP multi-physics and NLDAS multi-model ensembles. The average of the error correlations among the NLDAS models is less than 0.5 in most RFCs. They are independent models. This assessment based on error correlation is consistent with the assessment of model similarity by Kumar et al. (2017). However, the error correlation is high in NW, CN, and CB (greater than 0.5), which was not reported by Kumar et al. (2017). This may be partly caused by systematic errors in the forcing and evaluation data. The error stemming from the forcings can propagate into all the model outputs, yielding a high error correlation. Compared with the NLDAS models, the error correlation of the Noah-MP multi-physics ensemble is higher for both the annual cycle and interannual anomaly, suggesting that the Noah-MP ensemble members are not adequately independent. This high error correlation corresponds well to the low ensemble skill gain shown in Table 3.

4.4. Asymmetric responses of the PEM to the inclusion of the most versus least independent members

The above analyses show that there is a correspondence between the low ensemble skill gain and low independence. In this section, we further confirm the effects of independence by removing the most and least independent members from consideration. The most independent member has the lowest average error correlation with the other members, whereas the least independent member has the highest average value. Figure 8 shows the performance of the

Noah-MP multi-physics ensemble mean at the 12 RFCs. The immediate left points from the center denote the cases in which the least independent one or five members are removed, whereas the immediate right points denote the cases in which the most independent one or five members are removed. Note that independence is measured by the average of the error correlation. Figure 8 also shows the responses of the PEM to the removal of the worst-performing one or five members (left) and the best-performing one or five members (right). The effects of independence and constituent performance can thus be compared.

Figure 8 highlights four interesting observations. First, there is an asymmetry in the response of the PEM to the removal of the best- versus worst-performing members, which are consistent with previous studies on hydrological ensemble simulations (Ajami et al., 2006). Second, removing the most independent members degrades the performance of the ensemble mean, suggesting that the superiority of the ensemble mean could come from these independent members. Third, removing the least independent members does not have an obvious impact on the PEM, suggesting that the Noah-MP multi-physics ensemble can be optimized by eliminating these ensemble members. Fourth, the asymmetry of the independence effects is more pronounced than the asymmetry of the performance effects.

5 Conclusions

In this study, we compared the Noah-MP multi-physics ensemble and the NLDAS multi-model ensemble in simulating the annual cycle and interannual variability of runoff at 12 RFCs of the CONUS. The performance of the constituent members, the best member, the PEM, and its relation to inter-member independence are analyzed. The results are summarized below.

First, on average, the 48 Noah-MP configurations outperform the NLDAS models. In snow-dominated regions of the CONUS, Noah-MP can better capture the timing of the spring runoff peak from snowmelt. The models scatter more in arid than in humid RFCs, and the spread in arid RFCs is manifested in the different variability. The Noah-MP configurations with groundwater dynamics produce a distinguishable correlation coefficient to those without groundwater. The difference is more pronounced for the interannual anomaly than for the annual cycle.

Second, the Noah-MP ensemble contains the best-performing member among all the constituents of the two ensembles. This is the result of the broad coverage of the feasible model physics space. However, the best member varies significantly with the RFCs and differs between the annual cycle and interannual anomaly.

Third, the arithmetic average of the NLDAS models shows comparable performance to the Noah-MP multi-physics ensemble mean. This hints that the ensemble skill gain obtained from the Noah-MP multi-physics ensemble is significantly low compared with that obtained from the NLDAS multi-model ensemble. The low ensemble skill gain corresponds well to the high error correlation among the ensemble members.

Fourth, there is an asymmetry in the responses of the PEM to the inter-member independence. The performance of the ensemble mean deteriorates when the most independent members are removed, whereas it shows little change when the least independent members are removed. It is crucial for an ensemble to include independent members.

This study highlights the importance of the independence among the constituent members of an ensemble. This issue has been overlooked in many previous studies on Bayesian model averaging (Ajami et al., 2007) and ensemble optimization (Gan et al., 2019). Its negative effect is pronounced for the multi-physics ensemble. As shown in equation (4), the PEM is the sum of the AEP and ensemble skill gain. Improved ensemble averaging methods that can consider both the performance and independence of an ensemble member are desirable. Improved ensemble optimization methods that can select a skillful and independent subset are also crucial for reducing the high computation cost of the multi-physics ensembles, which is crucial for continental-scale operational systems (e.g., NLDAS).

Acknowledgments and Data

We thank Dr. Ruiqiang Ding from the Institute of Atmospheric Physics, Chinese Academy of Sciences and Dr. Dejian Yang from Hohai University for providing the helpful information and comments on this work. This work was supported by the National Key Research and Development Program of China (2018YFA0606004) and the National Natural Science Foundation of China (91337217, 41375088, and 41605062). The NLDAS-2 static data and meteorological forcing data were obtained from Goddard Earth Sciences Data and Information Services Center (https://disc.sci.gsfc.nasa.gov/uui/datasets?keywords=NLDAS). The USGS HUC8 runoff data were downloaded from the USGS Water Watch website (https://waterwatch.usgs.gov/?id=romap3). The model outputs were generated from the Noah-MP version 3.6, which can be downloaded from the website of Research and Applications Laboratory at the National Center for Atmospheric Research (https://ral.ucar.edu/solutions/products/noah-multiparameterization-land-surface-model-noah-mp-lsm).

References

Ajami, N. K., Duan, Q. Y., Gao, X. G., & Sorooshian, S. (2006). Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model

Intercomparison Project results. *Journal of Hydrometeorology, 7*(4), 755-768. https://doi.org/10.1175/jhm519.1

Ajami, N. K., Duan, Q. Y., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research, 43*(1), W01403. https://doi.org/10.1029/2005wr004745

Allen, G. H., David, C. H., Andreadis, K. M., Hossain, F., & Famiglietti, J. S. (2018). Global Estimates of River Flow Wave Travel Times and Implications for Low-Latency Satellite Data. *Geophysical Research Letters, 45*(15), 7551-7560. https://doi.org/10.1029/2018GL077914

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). FAO Irrigation and drainage paper 56. In *Crop evapotranspiration - Guidelines for computing crop water requirements* (pp. 300). Rome, Italy: Food and Agriculture Organization of the United Nations.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., & Martel, J. L. (2015). A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *Journal of Hydrology, 529*, 754-767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Barlage, M., Tewari, M., Chen, F., Miguez-Macho, G., Yang, Z. L., & Niu, G. Y. (2015). The effect of groundwater interaction in North American regional climate simulations with WRF/Noah-MP. *Climatic Change, 129*(3-4), 485-498. https://doi.org/10.1007/s10584-014-1308-8

Beck, H. E., van Dijk, A., de Roo, A., Dutra, E., Fink, G., Orth, R., & Schellekens, J. (2017). Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrology and Earth System Sciences, 21*(6), 2881-2903. https://doi.org/10.5194/hess-21-2881-2017

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Menard, C. B., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description - Part 1: Energy and water fluxes. *Geoscientific Model Development, 4*(1), 595-640. https://doi.org/10.5194/gmd-4-677-2011

Bohn, T. J., Sonessa, M. Y., & Lettenmaier, D. P. (2010). Seasonal Hydrologic Forecasting: Do Multimodel Ensemble Averages Always Yield Improvements in Forecast Skill? *Journal of Hydrometeorology, 11*(6), 1358-1372. https://doi.org/10.1175/2010jhm1267.1

Budyko, M. I. (1974). *Climate and Life*. New York: Academic Press.

Cai, X. T., Yang, Z. L., David, C. H., Niu, G. Y., & Rodell, M. (2014a). Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin. *Journal of Geophysical Research-Atmospheres, 119*(1), 23-38. https://doi.org/10.1002/2013jd020792

Cai, X. T., Yang, Z. L., Xia, Y. L., Huang, M. Y., Wei, H. L., Leung, L. R., & Ek, M. B. (2014b). Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS test bed. *Journal of Geophysical Research-Atmospheres, 119*(24), 13751-13770. https://doi.org/10.1002/2014JD022113

Chen, F., Mitchell, K., Schaake, J., Xue, Y. K., Pan, H. L., Koren, V., et al. (1996). Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research-Atmospheres, 101*(D3), 7251-7268. https://doi.org/10.1029/95JD02165

Cherkauer, K. A., Bowling, L. C., & Lettenmaier, D. P. (2003). Variable infiltration capacity cold land process model updates. *Global and Planetary Change, 38*(1-2), 151-159. https://doi.org/10.1016/s0921-8181(03)00025-0

Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description - Part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, *4*(3), 701-722. https://doi.org/10.5194/gmd-4-701-2011

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research, 47*(9), W09301. https://doi.org/10.1029/2010wr009827

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D., & Woods, R. A. (2010). Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses. *Hydrological Processes, 25*(4), 523-543. https://doi.org/10.1002/hyp.7902

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015a). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research, 51*(4), 2498-2514. https://doi.org/10.1002/2015wr017198

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015b). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research, 51*(4), 2515-2542. https://doi.org/10.1002/2015wr017200

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research, 52*(3), 2350-2365. https://doi.org/10.1002/2015wr017910

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., et al. (2003). Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research-Atmospheres, 108*(D22), 8842. https://doi.org/10.1029/2002jd003118

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes, 28*(25), 6135-6150. https://doi.org/10.1002/hyp.10096

Dickinson, R. E., Shaikh, M., Bryant, R., & Graumlich, L. (1998). Interactive canopies for a climate model. *Journal of Climate, 11*(11), 2823-2836. https://doi.org/10.1175/1520-0442(1998)011<2823:ICFACM>2.0.CO;2

Douinot, A., Roux, H., Garambois, P. A., & Dartus, D. (2018). Using a multi-hypothesis framework to improve the understanding of flow dynamics during flash floods. *Hydrology and Earth System Sciences*, *22*(10), 5317-5340. https://doi.org/10.5194/hess-22-5317-2018

Duan, Q. Y., Ajami, N. K., Gao, X. G., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources, 30*(5), 1371-1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Ducharne, A., Koster, R. D., Suarez, M. J., & Kumar, P. (1999). A catchment-based land surface model for GCMs and the framework for its evaluation. *Physics and Chemistry of the*

Earth Part B-Hydrology Oceans and Atmosphere, 24(7), 769-773. https://doi.org/10.1016/s1464-1909(99)00078-7

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., et al. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research-Atmospheres, 108*(D22), 8851. https://doi.org/10.1029/2002JD003296

Gan, Y. J., Liang, X. Z., Duan, Q. Y., Chen, F., Li, J. D., & Zhang, Y. (2019). Assessment and Reduction of the Physical Parameterization Uncertainty for Noah-MP Land Surface Model. *Water Resources Research, 55*(7), 5518-5538. https://doi.org/10.1029/2019wr024814

Gao, X., & Dirmeyer, P. A. (2006). A multimodel analysis, validation, and transferability study of global soil wetness products. *Journal of Hydrometeorology*, *7*(6), 1218-1236. https://doi.org/10.1175/jhm551.1

Georgakakos, K. P., Seo, D. J., Gupta, H., Schaake, J., & Butts, M. B. (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology, 298*(1-4), 222-241. https://doi.org/10.1016/j.jhydrol.2004.03.037

Goodwell, A. E., Jiang, P. S., Ruddell, B. L., & Kumar, P. (2020). Debates-Does Information Theory Provide a New Paradigm for Earth Science? Causality, Interaction, and Feedback. *Water Resources Research, 56*(2), 12. https://doi.org/10.1029/2019wr024940

Goodwell, A. E., & Kumar, P. (2017a). Temporal Information Partitioning Networks (TIPNets): A process network approach to infer ecohydrologic shifts. *Water Resources Research, 53*(7), 5899-5919. https://doi.org/10.1002/2016wr020218

Goodwell, A. E., & Kumar, P. (2017b). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, *53*(7), 5920-5942. https://doi.org/10.1002/2016wr020216

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., et al. (2012a). Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe. *Journal Of Hydrometeorology, 13*(2), 604-620. https://doi.org/10.1175/jhm-d-11-083.1

Gudmundsson, L., Wagener, T., Tallaksen, L. M., & Engeland, K. (2012b). Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Water Resources Research, 48*(11), W11504. https://doi.org/10.1029/2011wr010911

Guo, Z. C., Dirmeyer, P. A., Gao, X., & Zhao, M. (2007). Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Quarterly Journal of the Royal Meteorological Society*, *133*(624), 731-747. https://doi.org/10.1002/qj.48

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1-2), 80-91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M., & Andreassian, V. (2014). Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences, 18*(2), 463-477. https://doi.org/10.5194/hess-18-463-2014

Gutman, G., & Ignatov, A. (1998). The derivation of the green vegetation fraction from NOAA/AVHRR data for use in numerical weather prediction models. *International Journal of Remote Sensing, 19*(8), 1533-1543. https://doi.org/10.1080/014311698215333

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus Series a-Dynamic Meteorology and Oceanography, 57*(3), 219-233. https://doi.org/10.1111/j.1600-0870.2005.00103.x

Hsu, K. L., Moradkhani, H., & Sorooshian, S. (2009). A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research, 45*(12), W00B12. https://doi.org/10.1029/2008wr006824

Koren, V., Schaake, J., Mitchell, K., Duan, Q. Y., Chen, F., & Baker, J. M. (1999). A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. *Journal of Geophysical Research-Atmospheres, 104*(D16), 19569-19585. https://doi.org/10.1029/1999JD900232

Koster, R. D., & Suarez, M. J. (1992). Modeling the land surface boundary in climate models as a composite of independent vegetation stands. *Journal of Geophysical Research-Atmospheres, 97*(D3), 2697-2715. https://doi.org/10.1029/91JD01696

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., et al. (2010). Ensemble evaluation of hydrological model hypotheses. *Water Resources Research, 46*(7), W07516. https://doi.org/10.1029/2009wr007845

Kumar, S. V., Wang, S. G., Mocko, D. M., Peters-Lidard, C. D., & Xia, Y. L. (2017). Similarity Assessment of Land Surface Model Outputs in the North American Land Data Assimilation System. *Water Resources Research, 53*(11), 8941-8965. https://doi.org/10.1002/2017wr020635

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., et al. (2019). The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *Journal of Advances in Modeling Earth Systems, 11*(12), 4245-4287. https://doi.org/10.1029/2018ms001583

Li, J. D., Zhang, G., Chen, F., Peng, X. D., & Gan, Y. J. (2019). Evaluation of Land Surface Subprocesses and Their Impacts on Model Performance With Global Flux Data. *Journal of Advances in Modeling Earth Systems, 11*(5), 1329-1348. https://doi.org/10.1029/2018ms001606

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research-Atmospheres, 99*(D7), 14415-14428. https://doi.org/10.1029/94JD00483

Livneh, B., Xia, Y. L., Mitchell, K. E., Ek, M. B., & Lettenmaier, D. P. (2010). Noah LSM Snow Model Diagnostics and Enhancements. *Journal of Hydrometeorology, 11*(3), 721-738. https://doi.org/10.1175/2009jhm1174.1

Long, D., Longuevergne, L., & Scanlon, B. R. (2014). Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research, 50*(2), 1131-1151. https://doi.org/10.1002/2013wr014581

Ma, N., Niu, G. Y., Xia, Y. L., Cai, X. T., Zhang, Y. S., Ma, Y. M., & Fang, Y. H. (2017). A Systematic Evaluation of Noah-MP in Simulating Land-Atmosphere Energy, Water, and Carbon Exchanges Over the Continental United States. *Journal of Geophysical Research-Atmospheres, 122*(22), 12245-12268. https://doi.org/10.1002/2017jd027597

Maidment, D. R. (2017). Conceptual Framework for the National Flood Interoperability Experiment. *Journal of the American Water Resources Association, 53*(2), 245-257. https://doi.org/10.1111/1752-1688.12474

Marshall, L., Sharma, A., & Nott, D. (2006). Modeling the catchment via mixtures: Issues of model specification and validation. *Water Resources Research*, *42*(11), W11409. https://doi.org/10.1029/2005wr004613

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., & Woods, R. A. (2010). Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure. *Hydrological Processes*, *25*(4), 511-522. https://doi.org/10.1002/hyp.7841

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research-Atmospheres*, *109*(D7), D07S90. https://doi.org/10.1029/2003jd003823

Monteith, J. L. (1965). Evaporation and environment. *Symposia of the Society for Experimental Biology*, *19*, 205-234.

Niu, G. Y., & Yang, Z. L. (2006). Effects of Frozen Soil on Snowmelt Runoff and Soil Water Storage at a Continental Scale. *Journal of Hydrometeorology*, *7*(5), 937-952. https://doi.org/10.1175/JHM538.1

Niu, G. Y., Yang, Z. L., Dickinson, R. E., Gulden, L. E., & Su, H. (2007). Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *Journal of Geophysical Research-Atmospheres*, *112*(D7), D07103. https://doi.org/10.1029/2006JD007522

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research-Atmospheres*, *116*, D12109. https://doi.org/10.1029/2010JD015139

Oudin, L., Andreassian, V., Mathevet, T., Perrin, C., & Michel, C. (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resources Research*, *42*(7), W07410. https://doi.org/10.1029/2005wr004636

Penman, H. L. (1956). Estimating evaporation. *Eos, Transactions American Geophysical Union*, *37*(1), 43-50. https://doi.org/10.1029/TR037I001P00043

Pinker, R. T., Tarpley, J. D., Laszlo, I., Mitchell, K. E., Houser, P. R., Wood, E. F., et al. (2003). Surface radiation budgets in support of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data Assimilation System (NLDAS) Project. *Journal of Geophysical Research-Atmospheres*, *108*(D22), 8844. https://doi.org/10.1029/2002jd003301

Sellers, P. J., Mintz, Y., Sud, Y. C., & Dalcher, A. (1986). A Simple Biosphere Model (SIB) for use within General Circulation Models *Journal of the Atmospheric Sciences*, *43*(6), 505-531. https://doi.org/10.1175/1520-0469(1986)043<0505:ASBMFU>2.0.CO;2

Shamseldin, A. Y., Oconnor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology*, *197*(1-4), 203-229. https://doi.org/10.1016/s0022-1694(96)03259-3

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. *Water Resources Research*, *55*(2), 1510-1530. https://doi.org/10.1029/2018wr023197

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research-Atmospheres, 106*(D7), 7183-7192. https://doi.org/10.1029/2000JD900719

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., & Robinson, B. A. (2006). Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Research Letters, 33*(19), L19817. https://doi.org/10.1029/2006gl027126

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research, 43*(1), W01411. https://doi.org/10.1029/2005wr004838

Wei, H. L., Xia, Y. L., Mitchell, K. E., & Ek, M. B. (2013). Improvement of the Noah land surface model for warm season processes: evaluation of water and energy flux simulation. *Hydrological Processes, 27*(2), 297-303. https://doi.org/10.1002/hyp.9214

Winter, C. L., & Nychka, D. (2009). Forecasting skill of model averages. *Stochastic Environmental Research and Risk Assessment, 24*(5), 633-638. https://doi.org/10.1007/s00477-009-0350-y

Xia, Y. L., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Brewer, M., et al. (2016a). Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research-Atmospheres, 121*(6), 2750-2779. https://doi.org/10.1002/2015jd023733

Xia, Y. L., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Kumar, S., et al. (2016b). Basin-scale assessment of the land surface energy budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research-Atmospheres, 121*(1), 196-220. https://doi.org/10.1002/2015JD023889

Xia, Y. L., Ek, M. B., Wu, Y. H., Ford, T., & Quiring, S. M. (2015). Comparison of NLDAS-2 Simulated and NASMD Observed Daily Soil Moisture. Part I: Comparison and Analysis. *Journal of Hydrometeorology, 16*(5), 1962-1980. https://doi.org/10.1175/jhm-d-14-0096.1

Xia, Y. L., Hao, Z. C., Shi, C. X., Li, Y. H., Meng, J. S., Xu, T. R., et al. (2019). Regional and Global Land Data Assimilation Systems: Innovations, Challenges, and Prospects. *Journal of Meteorological Research, 33*(2), 159-189. https://doi.org/10.1007/s13351-019-8172-4

Xia, Y. L., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L. F., et al. (2012a). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *Journal of Geophysical Research-Atmospheres, 117*(D3), D03110. https://doi.org/10.1029/2011jd016051

Xia, Y. L., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012b). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres, 117*(D3), D03109. https://doi.org/10.1029/2011JD016048

Xia, Y. L., Mocko, D., Huang, M. Y., Li, B. L., Rodell, M., Mitchell, K. E., et al. (2017). Comparison and Assessment of Three Advanced Land Surface Models in Simulating Terrestrial Water Storage Components over the United States. *Journal of Hydrometeorology, 18*(3), 625-649. https://doi.org/10.1175/jhm-d-16-0112.1

Yang, R. H., Cohn, S. E., da Silva, A., Joiner, J., & Houser, P. R. (2003). Tangent linear analysis of the Mosaic land surface model. *Journal of Geophysical Research-Atmospheres, 108*(D2), 4054. https://doi.org/10.1029/2002JD002410

Yang, Z. L., Niu, G. Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research: Atmospheres, 116*(D12), D12110. https://doi.org/10.1029/2010JD015140

Yoo, J. H., & Kang, I. S. (2005). Theoretical examination of a multi-model composite for seasonal prediction. *Geophysical Research Letters, 32*(18), L18707. https://doi.org/10.1029/2005gl023513

You, Y. H., Huang, C. L., Yang, Z. L., Zhang, Y., Bai, Y. L., & Gu, J. (2020). Assessing Noah-MP Parameterization Sensitivity and Uncertainty Interval Across Snow Climates. *Journal of Geophysical Research-Atmospheres, 125*(4), e2019JD030417. https://doi.org/10.1029/2019jd030417

Zhang, B. Q., Xia, Y. L., Long, B., Hobbins, M., Zhao, X. N., Hain, C., et al. (2020). Evaluation and comparison of multiple evapotranspiration data models over the contiguous United States: Implications for the next phase of NLDAS (NLDAS-Testbed) development. *Agricultural and Forest Meteorology*, *280*, 107810. https://doi.org/10.1016/j.agrformet.2019.107810

Zhang, G., Chen, F., & Gan, Y. J. (2016). Assessing uncertainties in the Noah-MP ensemble simulations of a cropland site during the Tibet Joint International Cooperation program field campaign. *Journal of Geophysical Research-Atmospheres, 121*(16), 9576-9596. https://doi.org/10.1002/2016jd024928

Zheng, H., & Yang, Z. L. (2016). Effects of soil-type datasets on regional terrestrial water cycle simulations under different climatic regimes. *Journal of Geophysical Research-Atmospheres, 121*(24), 14387-14402. https://doi.org/10.1002/2016jd025187

Zheng, H., Yang, Z. L., Lin, P. R., Wei, J. F., Wu, W. Y., Li, L. C., et al. (2019). On the Sensitivity of the Precipitation Partitioning Into Evapotranspiration and Runoff in Land Surface Parameterizations. *Water Resources Research, 55*(1), 95-111. https://doi.org/10.1029/2017WR022236

Table 1. Comparison of the average (AEP) and the median performance between the Noah-MP multi-physics and the NLDAS multi-model ensembles.

RFC	Annual cycle				Interannual anomaly			
	AEP (TSS)		The median of the ensemble performance (TSS)		AEP (TSS)		The median of the ensemble performance (TSS)	
	Noah-MP	NLDAS	Noah-MP	NLDAS	Noah-MP	NLDAS	Noah-MP	NLDAS
NE	0.96	0.88	0.96	0.92	0.92	0.91	0.92	0.93
MA	0.96	0.90	0.97	0.93	0.88	0.92	0.89	0.95
OH	0.91	0.89	0.93	0.96	0.81	0.88	0.82	0.94
LM	0.95	0.90	0.95	0.90	0.93	0.93	0.95	0.93
SE	0.90	0.91	0.92	0.90	0.94	0.94	0.94	0.94
NC	0.90	0.79	0.90	0.77	0.87	0.85	0.90	0.85
NW	0.88	0.89	0.89	0.88	0.96	0.94	0.96	0.93
AB	0.95	0.84	0.96	0.83	0.90	0.75	0.92	0.81
MB	0.90	0.72	0.95	0.68	0.88	0.70	0.90	0.69
WG	0.74	0.73	0.75	0.84	0.88	0.73	0.90	0.79
CN	0.74	0.69	0.75	0.65	0.94	0.88	0.96	0.87
CB	0.82	0.77	0.82	0.79	0.84	0.75	0.85	0.76
RFC mean	0.88	0.83	0.90	0.84	0.90	0.85	0.91	0.87

Table 2. The best member of the Noah-MP multi-physics and the NLDAS multi-model ensembles in simulating the annual cycle and interannual anomaly over each RFC. The best members of the two ensembles are shown in italic. The labels are r1 for SIMGM, r2 for SIMTOP, r3 for NOAHR, and r4 for BATS; b1 for NOAHB, b2 for CLM, b3 for SSiB; t1 for M-O, t2 for Chen97; c1 for Ball–Berry, c2 for Jarvis.

	Annual cycle					Interannual anomaly				
RFC	Noah-MP		NLDAS			Noah-MP		NLDAS		
	Best member	Best TSS	Best member	Best TSS	Best ranking	Best member	Best TSS	Best member	Best TSS	Best ranking
NE	*r1b1t1c1*	0.99	Mosaic	0.92	44	*r4b1t1c2*	0.95	VIC	0.94	4
MA	*r3b3t2c2*	0.98	VIC	0.97	23	*r4b1t2c1*	0.97	VIC	0.96	6
OH	*r4b2t2c1*	1.00	VIC	1.00	5	r4b1t2c1	0.94	*VIC*	0.97	1
LM	*r2b1t1c1*	0.99	Mosaic	0.91	43	*r4b3t1c1*	0.99	Noah	0.96	24
SE	*r2b1t2c1*	0.99	VIC	0.98	11	*r2b1t2c1*	0.98	Mosaic	0.97	7
NC	*r4b1t1c1*	0.98	VIC	0.97	9	*r4b2t1c1*	0.96	VIC	0.93	18
NW	*r1b1t1c2*	0.96	*VIC*	0.96	1	*r3b2t1c1*	0.97	Noah	0.94	37
AB	*r2b1t2c1*	0.99	VIC	0.89	46	*r2b1t2c1*	0.95	Mosaic	0.82	44
MB	*r3b3t1c2*	0.99	Mosaic	0.90	31	*r2b1t2c1*	0.95	Mosaic	0.78	44
WG	*r1b3t1c2*	0.93	Noah	0.85	10	*r2b3t1c1*	0.95	Mosaic	0.90	24
CN	*r1b1t1c2*	0.91	Mosaic	0.86	7	*r2b1t1c2*	0.98	Mosaic	0.95	28
CB	*r1b3t1c1*	0.98	VIC	0.87	10	*r1b1t1c1*	0.91	VIC	0.87	19

Table 3. The performance of the ensemble mean (PEM) obtained from the Noah-MP multi-physics and NLDAS multi-model ensembles. Note that the ensemble mean is ranked within each of the two ensembles.

RFC	Annual cycle						Interannual anomaly					
	PEM (TSS)		Ensemble skill gain (TSS)		Ranking of the ensemble mean		PEM (TSS)		Ensemble skill gain (TSS)		Ranking of the ensemble mean	
	Noah-MP	NLDAS	Noah-MP	NLDAS	Noah-MP	NLDAS	Noah-MP	NLDAS	Noah-MP	NLDAS	Noah-MP	NLDAS
NE	0.98	0.92	0.02	0.04	12	1	0.94	0.96	0.03	0.05	5	0
MA	0.98	0.92	0.02	0.03	2	1	0.90	0.96	0.02	0.04	23	0
OH	0.93	0.94	0.02	0.06	23	2	0.82	0.95	0.01	0.06	28	0
LM	0.98	0.98	0.03	0.08	5	1	0.96	0.98	0.03	0.05	22	0
SE	0.95	0.98	0.05	0.08	12	1	0.97	0.97	0.04	0.03	3	0
NC	0.93	0.92	0.03	0.13	12	1	0.90	0.96	0.03	0.11	24	0
NW	0.90	0.92	0.01	0.03	22	0	0.97	0.96	0.01	0.03	15	0
AB	0.99	0.98	0.04	0.13	0	1	0.96	0.88	0.05	0.13	0	2
MB	0.96	0.80	0.06	0.08	18	0	0.94	0.85	0.07	0.15	3	2
WG	0.77	0.86	0.03	0.13	23	0	0.93	0.79	0.05	0.07	9	1
CN	0.74	0.69	0.00	0.00	26	0	0.96	0.90	0.01	0.01	24	3
CB	0.88	0.79	0.06	0.02	8	1	0.89	0.83	0.04	0.09	8	2
RFC mean	0.92	0.89	0.03	0.07	14	1	0.93	0.92	0.03	0.07	14	1

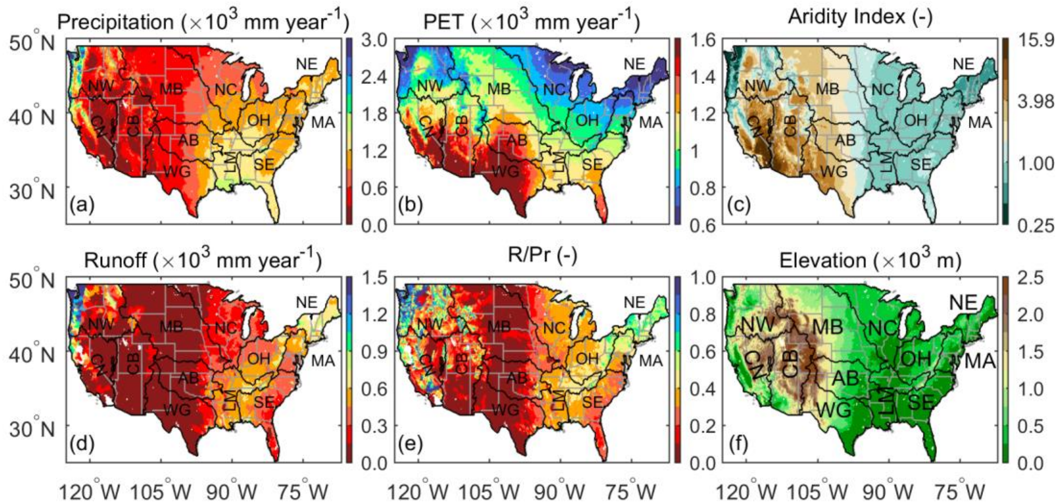


Figure 1. Spatial patterns of (a) multi-year averaged precipitation (1982 to 2011), (b) potential evapotranspiration, (c) Budyko's aridity index, (d) runoff, (e) runoff ratio (*R*/*P*), and (f) elevation over the CONUS. The RFC labels are NE for Northeast, MA for Mid-Atlantic, OH for Ohio, LM for Lower Mississippi, SE for Southeast, NC for North Central, NW for Northwest, AB for Arkansas, MB for Missouri, WG for West Gulf, CN for California-Nevada, and CB for Colorado.

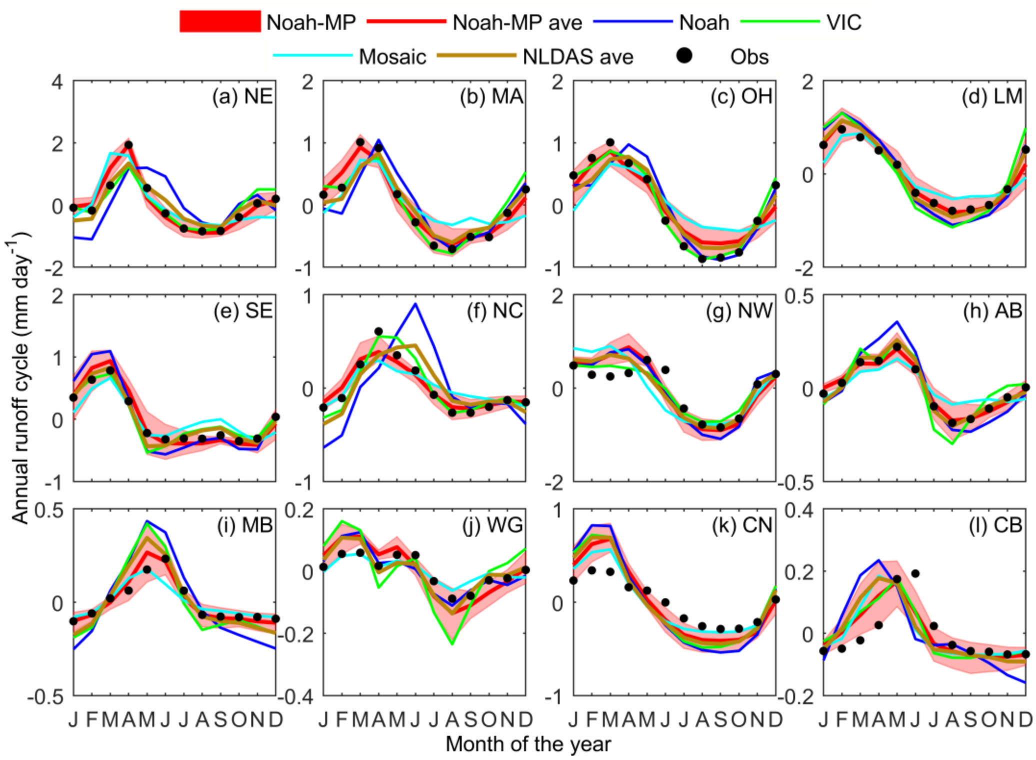


Figure 2. Modeled and observed annual cycle at each RFC. Black dots denote the observations. The shaded areas denote the maxima and minima of the 48-member Noah-MP ensemble. The solid red line denotes the Noah-MP multi-physics ensemble mean. The three NLDAS models (Noah, Mosaic, VIC) and their ensemble mean are denoted by the blue, green, cyan, and dark golden lines, respectively. The aridity of the 12 RFCs increases monotonically from the top to bottom, left to right (i.e., the most humid RFC on the top left, and the driest RFC on the bottom right).

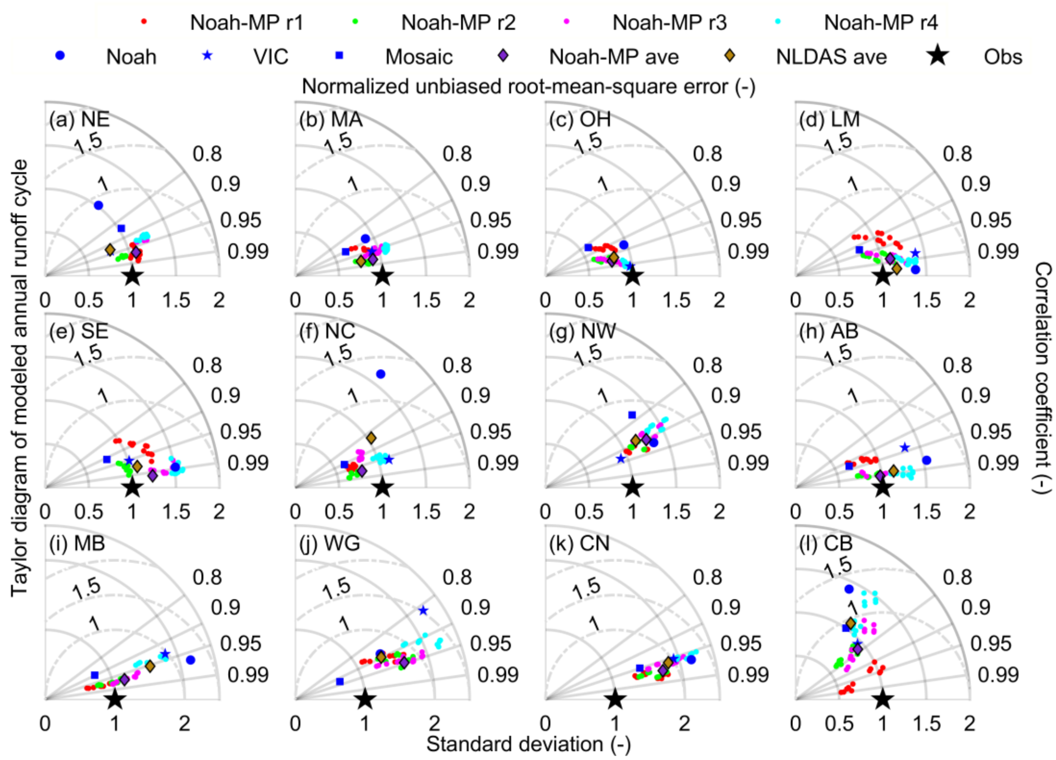


Figure 3. Normalized Taylor diagrams showing the performance of the modeled annual runoff cycle from the 48 Noah-MP multi-physics ensemble members, which are denoted by the red, green, magenta, and cyan dots for different runoff parameterizations, including SIMGM (r1), SIMTOP (r2), NOAHR (r3), and BATS (r4), the three NLDAS models (Noah, VIC, and Mosaic) (blue dot, star, and square), the Noah-MP multi-physics ensemble mean (purple diamond), and the NLDAS multi-model ensemble mean (dark golden diamond) at each RFC. The black star denotes the observations. The distance between a point of the model simulation to the observations denotes the normalized unbiased root-mean-square error. The radical lines denote the correlation coefficient, while the distance to the origin along the line denotes the normalized variability.

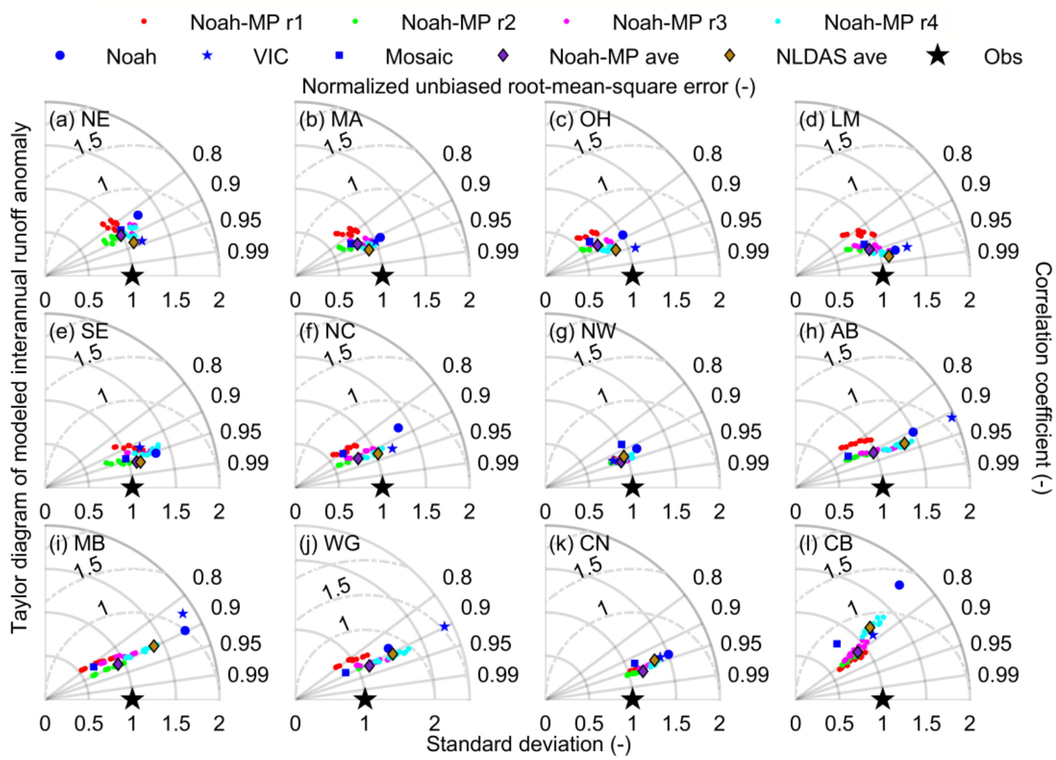


Figure 4. Same as Figure 3, but for the interannual anomaly.

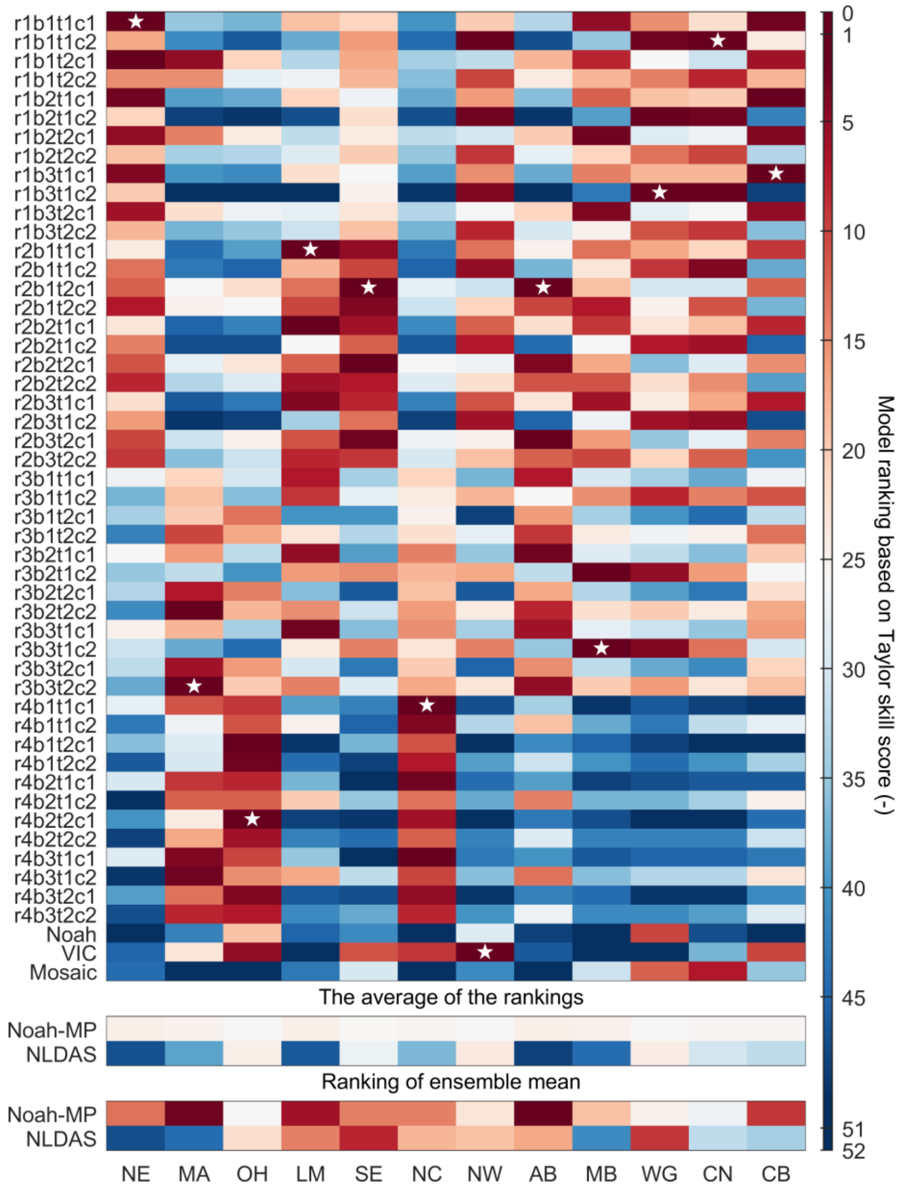
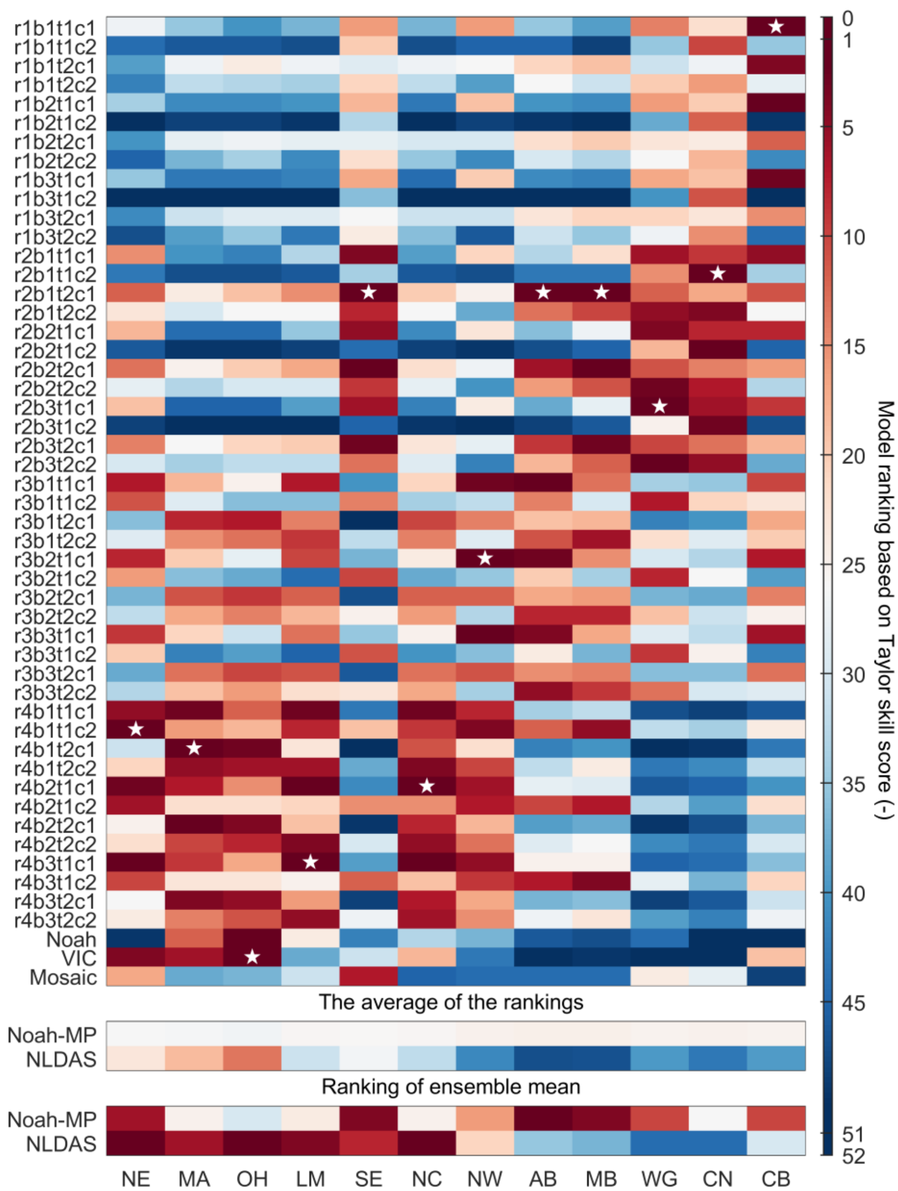


Figure 5. The performance ranking of the 48 Noah-MP multi-physics plus and the three NLDAS multi-model (Noah, VIC, and Mosaic) ensemble members for simulating the annual runoff cycle over the 12 RFCs. The ranking is across the two ensembles and ranges from 1 to 51 (51 = 48 + 3), from the best to the worst. Stars mark the best member among the two ensembles at each RFC. White indicates the performance median (a ranking of 26). Red indicates better than the median and blue indicates worse than the median. The bottom panel shows the ranking of the

ensemble mean, which ranges from 0 to 52. A ranking of 0 (52) indicates that it outperforms (underperforms) all the 51 constituent members. The labels are r1 for SIMGM, r2 for SIMTOP, r3 for NOAHR, and r4 for BATS; b1 for NOAHB, b2 for CLM, b3 for SSiB; t1 for M-O, t2 for Chen97; c1 for Ball–Berry, c2 for Jarvis.



893

894 **Figure 6**. Same as Figure 6, but for the interannual runoff anomaly.

895

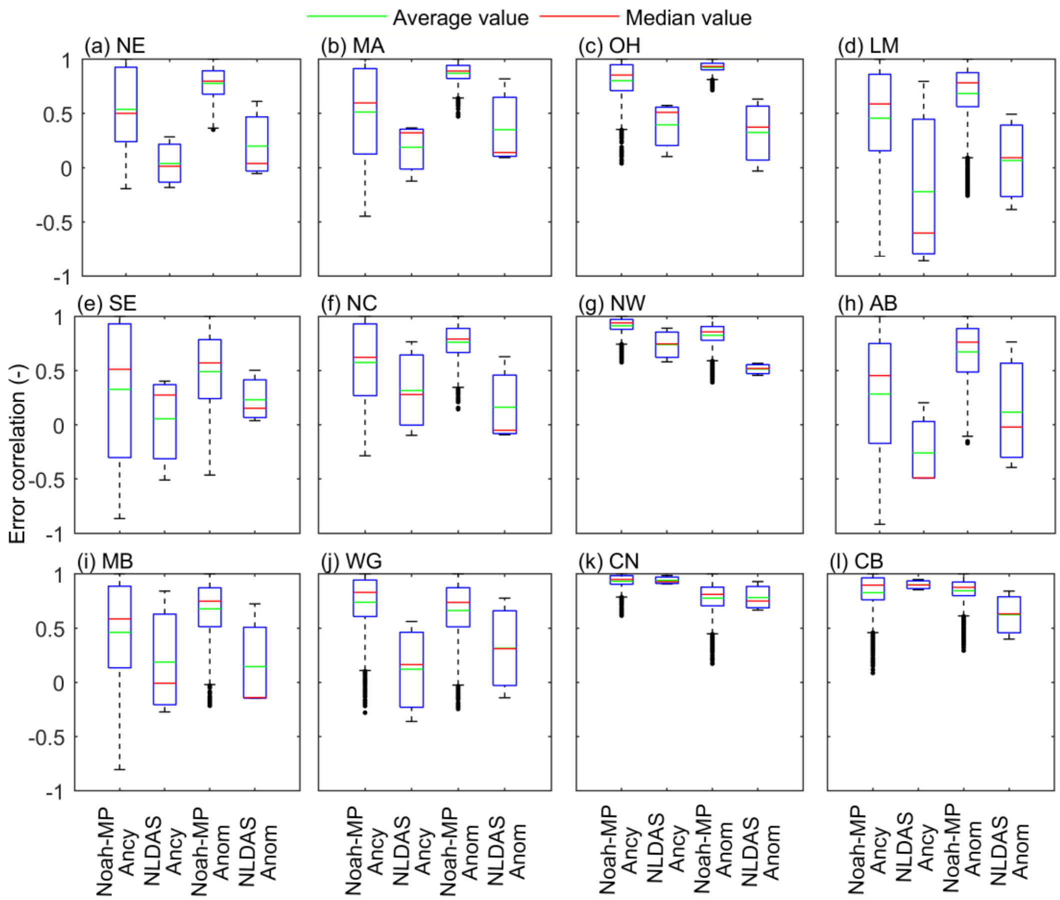


Figure 7. The error correlation in simulating the annual runoff cycle (Ancy) and interannual anomaly (Anom) for the Noah-MP multi-physics and NLDAS multi-model ensembles. There are 1128 ($48 \times 47/2$) samples for the Noah-MP multi-physics ensemble and 3 ($3 \times 2/2$) samples for the NLDAS multi-model ensemble. The upper and lower quantile lines show the 75th percentile value and 25th percentile value, respectively. The green and red lines denote the average and median values, respectively. The black dots denote the outliers, which are outside 1.5 times the interquartile range (the upper quantile value minus the lower quantile value) above the upper quartile and below the lower quartile.

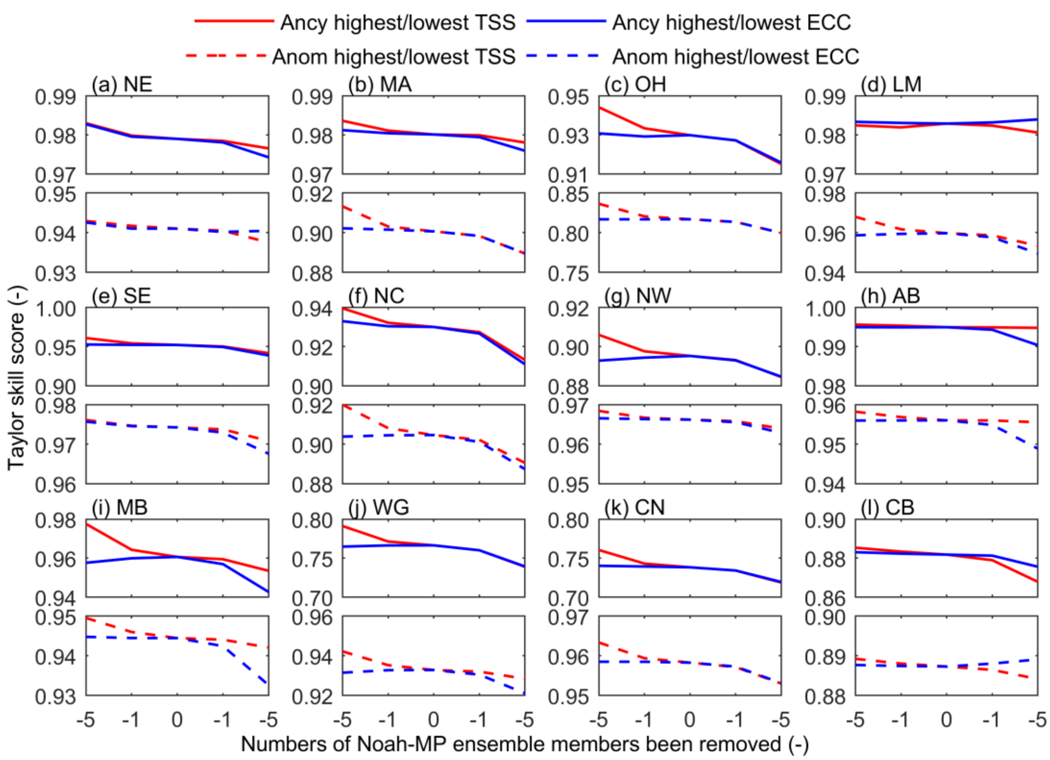
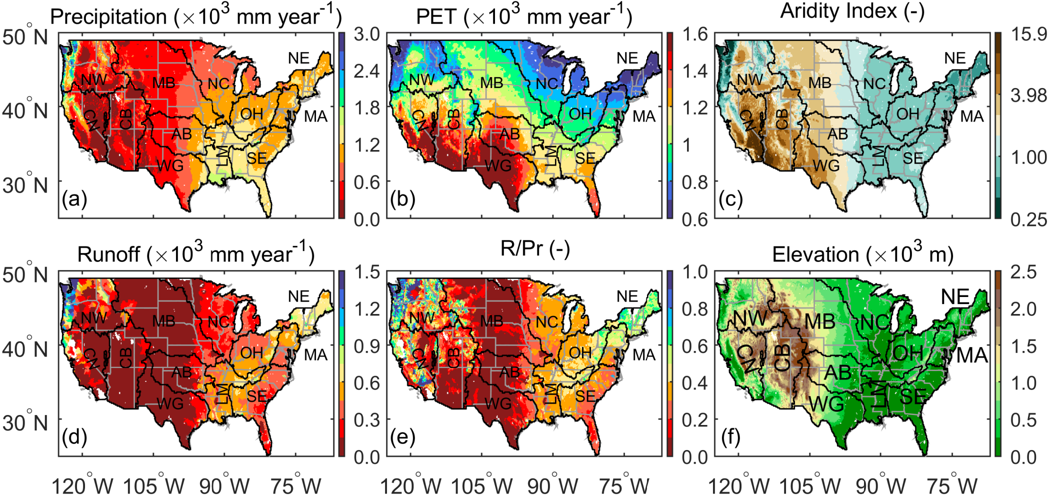


Figure 8. The performance of the Noah-MP multi-physics ensemble mean in simulating the annual runoff cycle (Ancy, upper subpanel, solid line) and interannual anomaly (Anom, lower subpanel, dash line) at the 12 RFCs. In each panel, the least independent or the worst-performing one or five member(s) are removed on the left, whereas the most independent or the best-performing one or five member(s) are removed on the right. The zero indicates the original Noah-MP ensemble with 48 members. The effect of removing the worst/best-performing (highest/lowest TSS) members is shown by the red lines and the effect of removing the least/most independent (highest/lowest ECC) members is shown by the blue lines.

Figure 1.



Precipitation (×10^3 mm year^-1)
PET (×10^3 mm year^-1)
Aridity Index (-)
Runoff (×10^3 mm year^-1)
R/Pr (-)
Elevation (×10^3 m)
50°N
40°N
30°N
120°W
105°W
90°W
75°W
NW
MB
NC
NE
OH
MA
CV
CB
AB
WG
LM
SE
(a)
(b)
(c)
(d)
(e)
(f)
3.0
2.4
1.8
1.2
0.6
0.0
1.6
1.4
1.2
1
0.8
0.6
15.9
3.98
1.00
0.25
1.5
1.2
0.9
0.6
0.3
0.0
1.0
0.8
0.6
0.4
0.2
0.0
2.5
2.0
1.5
1.0
0.5
0.0

Figure 2.

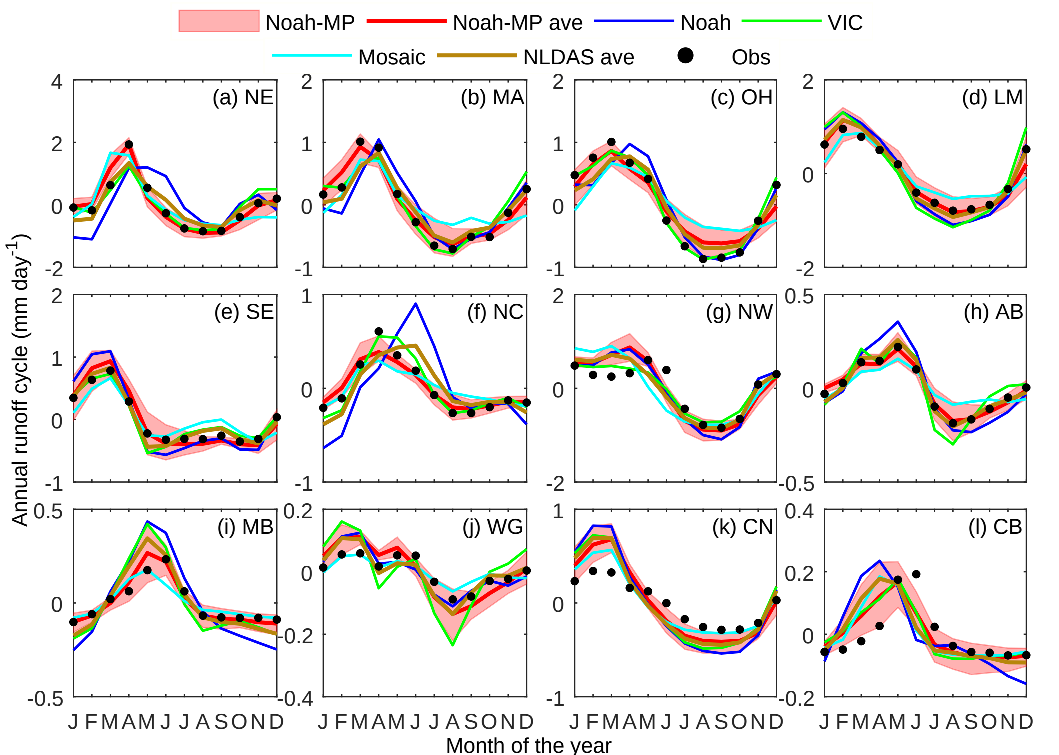
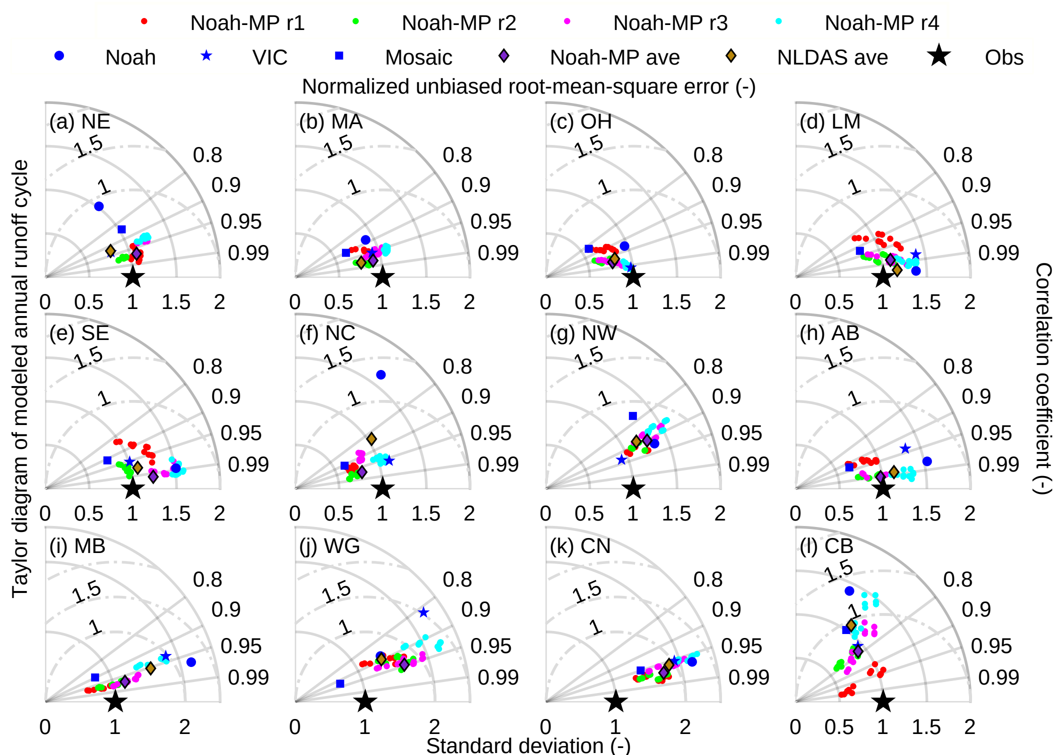

Noah-MP
Noah-MP ave
Noah
VIC
Mosaic
NLDAS ave
Obs
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB
Annual runoff cycle (mm day-1)
J F M A M J J A S O N D
Month of the year

Figure 3.



Noah-MP r1
Noah-MP r2
Noah-MP r3
Noah-MP r4
Noah
VIC
Mosaic
Noah-MP ave
NLDAS ave
Obs
Normalized unbiased root-mean-square error (-)
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB
Taylor diagram of modeled annual runoff cycle
Correlation coefficient (-)
Standard deviation (-)

Figure 4.

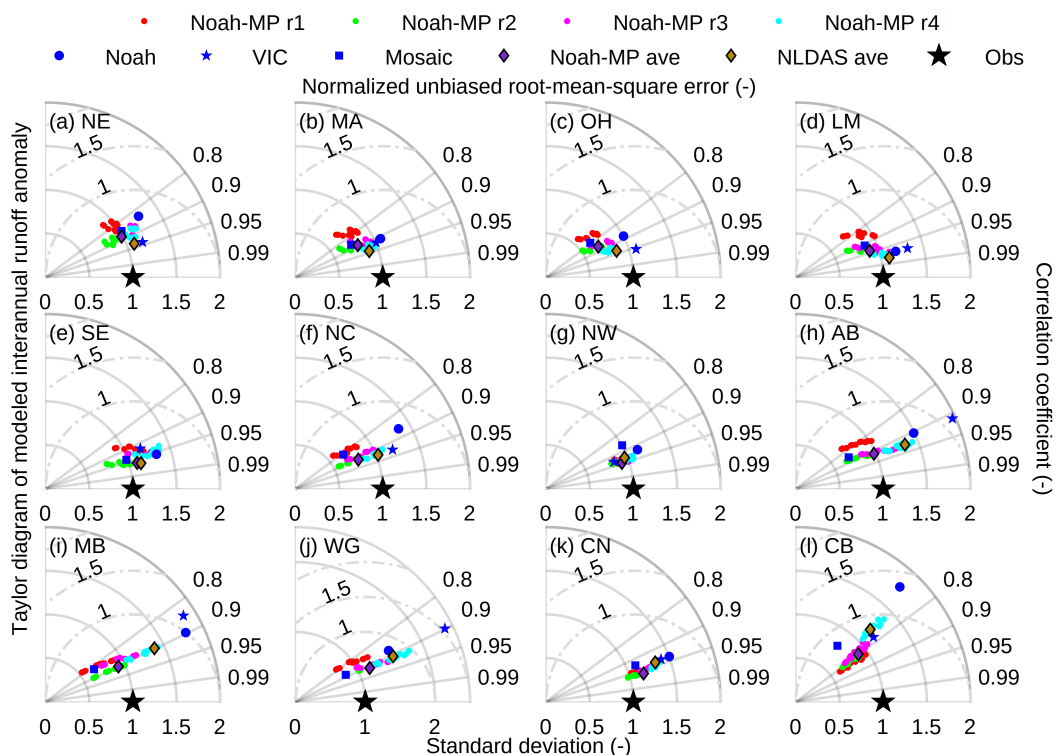

Noah-MP r1
Noah-MP r2
Noah-MP r3
Noah-MP r4
Noah
VIC
Mosaic
Noah-MP ave
NLDAS ave
Obs
Normalized unbiased root-mean-square error (-)
Taylor diagram of modeled interannual runoff anomaly
Correlation coefficient (-)
Standard deviation (-)
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB
0.8
0.9
0.95
0.99
1.5
1
0
0.5
1
1.5
2

Figure 5.

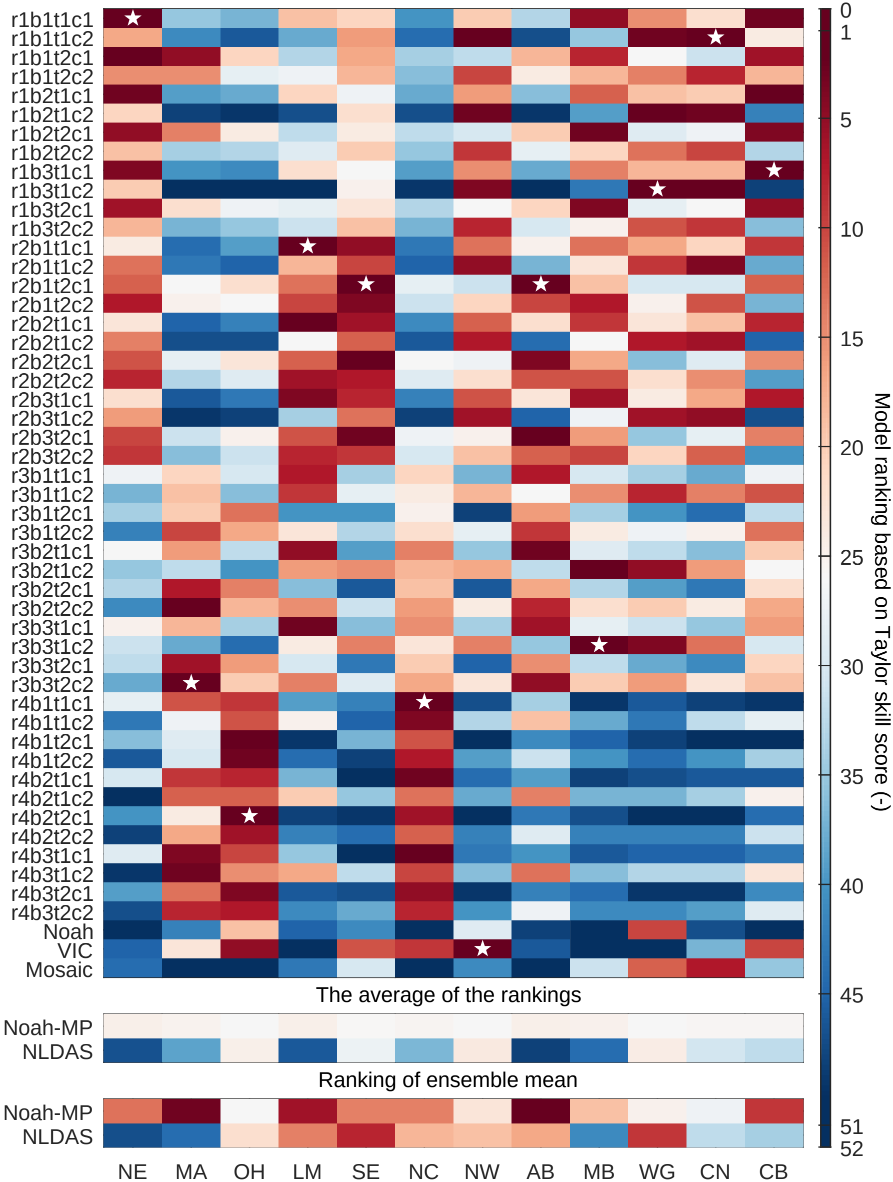

r1b1t1c1
r1b1t1c2
r1b1t2c1
r1b1t2c2
r1b2t1c1
r1b2t1c2
r1b2t2c1
r1b2t2c2
r1b3t1c1
r1b3t1c2
r1b3t2c1
r1b3t2c2
r2b1t1c1
r2b1t1c2
r2b1t2c1
r2b1t2c2
r2b2t1c1
r2b2t1c2
r2b2t2c1
r2b2t2c2
r2b3t1c1
r2b3t1c2
r2b3t2c1
r2b3t2c2
r3b1t1c1
r3b1t1c2
r3b1t2c1
r3b1t2c2
r3b2t1c1
r3b2t1c2
r3b2t2c1
r3b2t2c2
r3b3t1c1
r3b3t1c2
r3b3t2c1
r3b3t2c2
r4b1t1c1
r4b1t1c2
r4b1t2c1
r4b1t2c2
r4b2t1c1
r4b2t1c2
r4b2t2c1
r4b2t2c2
r4b3t1c1
r4b3t1c2
r4b3t2c1
r4b3t2c2
Noah
VIC
Mosaic
The average of the rankings
Noah-MP
NLDAS
Ranking of ensemble mean
Noah-MP
NLDAS
NE MA OH LM SE NC NW AB MB WG CN CB
0 1 5 10 15 20 25 30 35 40 45 51 52
Model ranking based on Taylor skill score (-)

Figure 6.

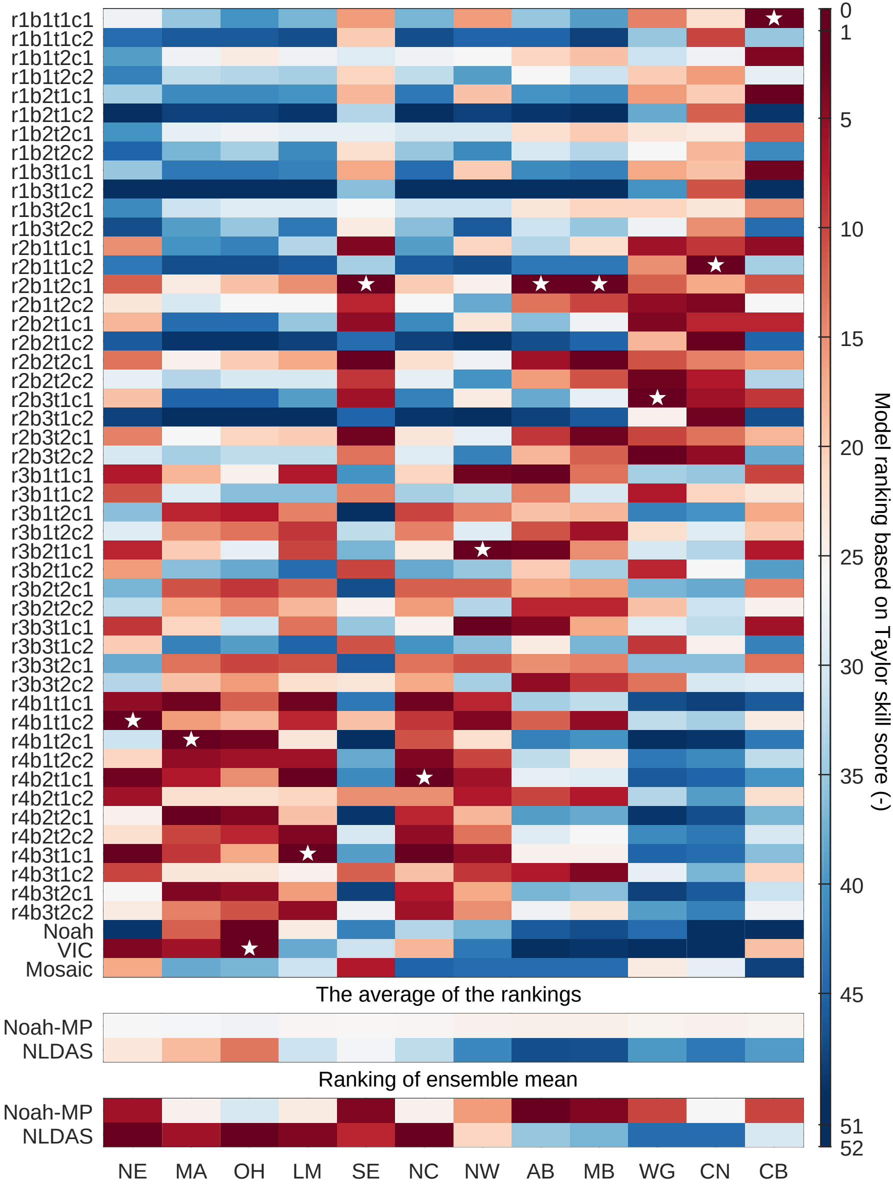
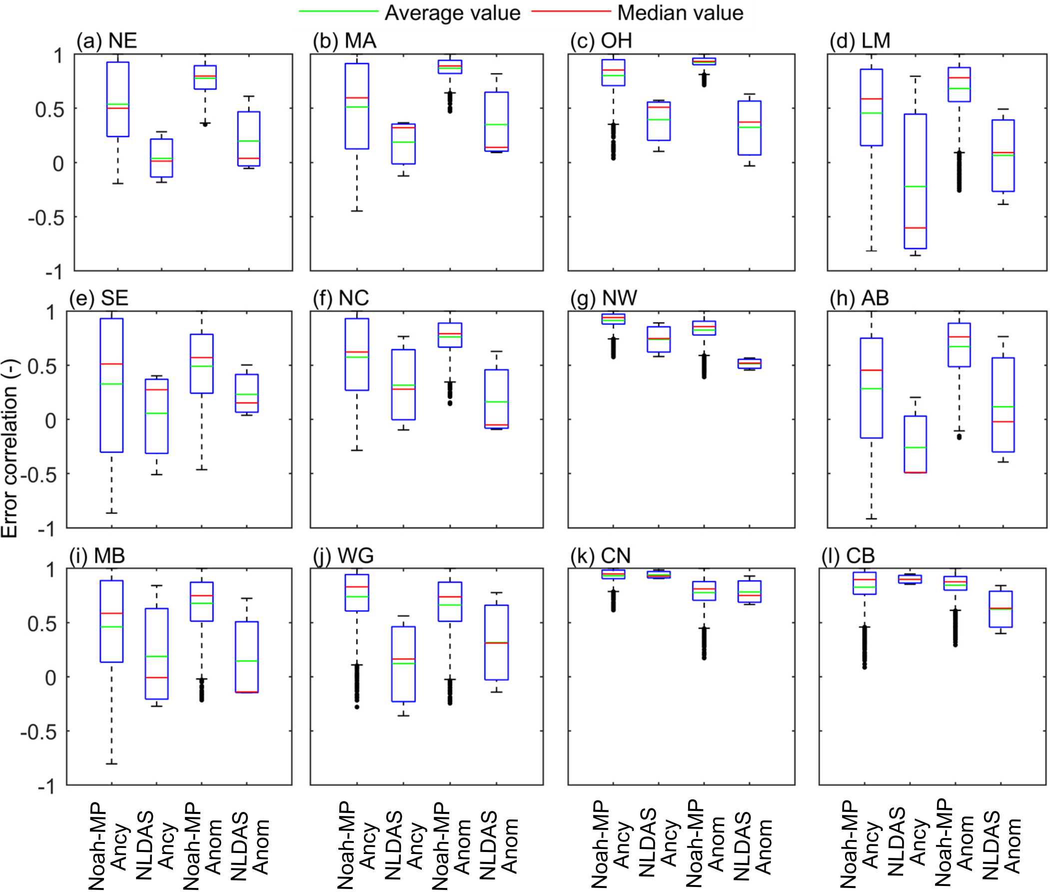
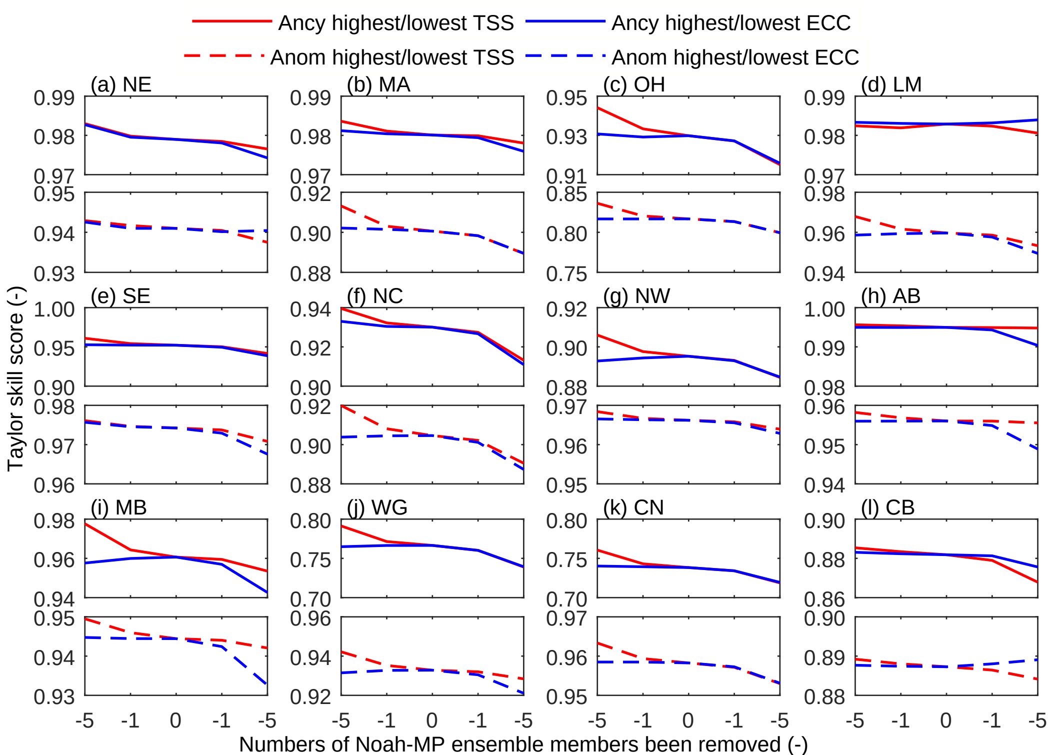

r1b1t1c1
r1b1t1c2
r1b1t2c1
r1b1t2c2
r1b2t1c1
r1b2t1c2
r1b2t2c1
r1b2t2c2
r1b3t1c1
r1b3t1c2
r1b3t2c1
r1b3t2c2
r2b1t1c1
r2b1t1c2
r2b1t2c1
r2b1t2c2
r2b2t1c1
r2b2t1c2
r2b2t2c1
r2b2t2c2
r2b3t1c1
r2b3t1c2
r2b3t2c1
r2b3t2c2
r3b1t1c1
r3b1t1c2
r3b1t2c1
r3b1t2c2
r3b2t1c1
r3b2t1c2
r3b2t2c1
r3b2t2c2
r3b3t1c1
r3b3t1c2
r3b3t2c1
r3b3t2c2
r4b1t1c1
r4b1t1c2
r4b1t2c1
r4b1t2c2
r4b2t1c1
r4b2t1c2
r4b2t2c1
r4b2t2c2
r4b3t1c1
r4b3t1c2
r4b3t2c1
r4b3t2c2
Noah
VIC
Mosaic
The average of the rankings
Noah-MP
NLDAS
Ranking of ensemble mean
Noah-MP
NLDAS
NE
MA
OH
LM
SE
NC
NW
AB
MB
WG
CN
CB
0
1
5
10
15
20
25
30
35
40
45
51
52
Model ranking based on Taylor skill score (-)

Figure 7.



Average value
Median value
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB
Error correlation (-)
1
0.5
0
-0.5
-1
Noah-MP Ancy
NLDAS Ancy
Noah-MP Anom
NLDAS Anom

Figure 8.


Ancy highest/lowest TSS
Ancy highest/lowest ECC
Anom highest/lowest TSS
Anom highest/lowest ECC
(a) NE
(b) MA
(c) OH
(d) LM
(e) SE
(f) NC
(g) NW
(h) AB
(i) MB
(j) WG
(k) CN
(l) CB
Taylor skill score (-)
Numbers of Noah-MP ensemble members been removed (-)