Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR024814

Assessment and Reduction of the Physical Parameterization Uncertainty for Noah-MP Land Surface Model

Yanjun Gan[1,2], Xin-Zhong Liang[3,4], Qingyun Duan[5], Fei Chen[1,6], Jianduo Li[1], and Yu Zhang[2]

[1]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China, [2]Department of Civil Engineering, University of Texas at Arlington, Arlington, TX, USA, [3]School of Atmospheric Science, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China, [4]Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA, [5]State Key Laboratory of Earth Surface Processes and Resource Ecology, Faulty of Geographical Science, Beijing Normal University, Beijing, China, [6]National Center for Atmosphere Research, Boulder, CO, USA

Abstract The community Noah land surface model with multiparameterization options (Noah-MP) provides a plethora of model configurations with varying complexity for land surface modeling. The practical application of this model requires a basic understanding of the relative abilities of its various parameterization configurations in representing spatiotemporal variability and hydrologic connectivity. We designed an ensemble of 288 experiments from multiparameterization schemes of six physical processes to assess and reduce the structural uncertainty for land surface modeling over 10 hydrologic regions in China for the period 2001–2010. The observed latent heat (LH) was well reproduced by the ensemble. Meanwhile, most experiments underestimated sensible heat (SH) throughout the year and overestimated the cold season but underestimated the warm season terrestrial water storage anomaly (TWSA). The sensitive processes and best-performing schemes varied not only with regions but also among variables. The LH and SH were most sensitive to runoff-groundwater (RUN), surface heat exchange coefficient (SFC), and radiation transfer (RAD). The TWSA was dominated by RUN and RAD while largely influenced by soil moisture factor for stomatal resistance (BTR) and frozen soil permeability (INF) over some limited regions. By contrast, supercooled liquid water (FRZ) had little influence on all variables. Our optimization for individual variables produced high mean Taylor skill scores that ranged from 0.95–0.99 for LH, 0.82–0.99 for SH, and 0.63–0.95 for TWSA depending on regions. The simultaneous optimization made trade-off among the three variables, which improved TWSA performance at the cost of reducing the skill for LH and SH over a few regions.

Key Points:
- We designed 288 experiments from multiparameterization schemes of six physical processes to assess parameterization uncertainty of Noah-MP
- Dominant physical processes for energy and water fluxes simulations over 10 regions in China were identified with sensitivity scores
- Best scheme combinations for corresponding simulations were selected with multiple comparisons of the parameterization schemes

Supporting Information:
- Supporting Information S1

Correspondence to:
Y. Gan, and X.-Z. Liang,
yjgan@cma.gov.cn;
xliang@umd.edu

Citation:
Gan, Y., Liang, X.-Z., Duan, Q., Chen, F., Li, J., & Zhang, Y. (2019). Assessment and reduction of the physical parameterization uncertainty for Noah-MP Land Surface Model. *Water Resources Research*, *55*, 5518–5538. https://doi.org/10.1029/2019WR024814

Received 21 JAN 2019
Accepted 6 JUN 2019
Accepted article online 20 JUN 2019
Published online 8 JUL 2019

1. Introduction

Land surface models (LSMs) play an important role in numerical weather prediction and climate projection, as they control the water, energy, and nutrient exchanges between the land surface and atmosphere (Dickinson et al., 2006; Gan et al., 2015; Liang et al., 2012; Pitman, 2003). They have evolved from simple bucket models representing water and energy balances (Manabe, 1969) to complex models representing an increasing number of interactions and feedbacks between physical, chemical, and biological processes (Bonan et al., 2011; Chen & Dudhia, 2001; Choi et al., 2013; Choi & Liang, 2010; Dai et al., 2003; Dickinson et al., 1986; Liang et al., 1994; Oleson et al., 2008; Sellers et al., 1997). The diversity of land surface modeling approaches motivates the development of unified modeling frameworks that provide multiple schemes for the same processes, such as the Noah land surface model with multiparameterization options (Noah-MP; Niu et al., 2011; Yang et al., 2011), Joint UK Land Environment Simulator (JULES; Best et al., 2011; Clark et al., 2011), and Structure for Unifying Multiple Modeling Alternatives (SUMMA; Clark et al., 2015a, 2015b). Among them, Noah-MP has received increasing attention due to its large improvement of the widely used baseline model Noah (Chen & Dudhia, 2001; Ek et al., 2003) and its coupling to the community Weather Research and Forecasting (WRF) model (Barlage et al., 2015), WRF-Hydro model (Gochis et al., 2013), and National Water Model (NWM; Cosgrove et al., 2017).

The flexibility in the selection of different schemes under the same modeling framework accommodates a plethora of physics configurations with varying complexity. The total number of combinations of alternative schemes for the 10 processes of the original Noah-MP is up to 4,608 (Gayler et al., 2014) and this number has been increasing due to the incorporation of new schemes for even more physical processes in the latest version. Noah-MP has been evaluated at global scales (Yang, Niu, et al., 2011), regional scales (Cai et al., 2014; Chen et al., 2015; Ma et al., 2017; Niu et al., 2011; Park & Park, 2016; Xia et al., 2017; Zheng & Yang, 2016), basin scales (Cai et al., 2014; Cuntz et al., 2016; Mendoza et al., 2015), and site scales (Arsenault et al., 2018; Cai et al., 2016; Gao et al., 2015; Pilotto et al., 2015; Zhang et al., 2016; Zheng et al., 2015). However, most researchers used a single, subjective set of scheme combination mainly based on their own experience or previous studies. In reality, a scheme combination that performs in one situation may not be appropriate for another due to the substantial spatial and temporal heterogeneities in surface soil and vegetation characteristics and their associated physical processes (Clark et al., 2011; Gayler et al., 2014; Mendoza et al., 2015).

A systematic and rigorous intercomparison of the multiple parameterizations would help to identify the dominant processes and appropriate configurations for representing varying hydroclimate, soil, and vegetation conditions (Bai et al., 2009; Best et al., 2015; Henderson-Sellers et al., 1993). This would further help to (1) reinforce our understanding of model behavior across different regions (Ma et al., 2017), (2) provide guidance for improving model physics (Wang et al., 2018), (3) conduct multiparameterization ensemble prediction (Zheng & Yang, 2016), and (4) perform parameter sensitivity analysis and optimization (Arsenault et al., 2018; Duan et al., 2017; Gan et al., 2014). Nevertheless, the intercomparison work is not only constrained by the computational cost of a large number of scheme combinations, but also with great uncertainty owing to the interactions among different processes.

Previous studies on Noah-MP parameterization intercomparison were limited to a few sites, a few scheme combinations, or short-term simulations due to the otherwise high computational cost of model simulations. Moreover, most research focused on the sensitivities of physical processes, whereas few tried to optimize the combinations of parameterization schemes. Besides, a systematic intercomparison of the Noah-MP parameterization schemes across different hydrologic regions over the continental China domain has not been made. For example, Yang, Niu et al. (2011) conducted 36 ensemble experiments to explore the influences of soil moisture factor for stomatal resistance, runoff and groundwater, leaf dynamics, and stomatal resistance in simulating land surface fluxes over global 50 largest river basins for the period 1983–2006. They concluded that runoff and groundwater are predominant over the other processes in controlling globally averaged soil moisture and evapotranspiration and their relationship. Afterward, Hong et al. (2014) employed the microgenetic algorithm (Krishnakumar, 1990) to identify the best scheme combinations of Noah-MP for evapotranspiration and runoff simulation over the Han River basin in South Korea for the period 2001–2003. They found that radiation transfer, surface exchange coefficient, and runoff and groundwater are the most important processes, but the scheme sets chosen to produce the most accurate simulations are very different between evapotranspiration and runoff. Gayler et al. (2014) designed 720 ensemble experiments to identify the most responsive processes for water and energy fluxes simulations during the 2009 vegetation period at two agricultural field sites with different soil and hydroclimate conditions in Southwest Germany. The authors observed that the way Noah-MP responds to the alternative scheme options is quite different among output variables and research sites. Zhang et al. (2016) conducted 1,152 ensemble experiments for sensible and latent heat simulation in 2008 at the Dali site in Southwest China. They reported that stomatal resistance, soil moisture factor for stomatal resistance, runoff and groundwater, and surface exchange coefficient are the most important factors.

The main purpose of this paper is to assess and reduce the uncertainties associated with the physical parameterizations of Noah-MP in simulating land surface fluxes over 10 large hydrologic regions in China with varying hydroclimate, soil, and vegetation conditions. We accomplish this by directly comparing the performances among different parameterization schemes and objectively analyzing their differences with statistical methods. Section 2 describes the Noah-MP model, evaluation data, and analysis approach. Section 3 presents the results and discussion. Section 4 concludes our main findings.

2. Model Description, Evaluation Data, and Analysis Approach

2.1. Noah-MP Physical Parameterization Options

Noah-MP is a new-generation LSM enhanced from the Noah model through the incorporation of improved physics and multiple parameterization options for different land-atmosphere interaction processes. The major advances in model physics and related theories can be found in Niu et al. (2011) and Yang et al. (2011). We adopted the Noah-MP v1.6 and selected the control configuration for the following six processes as suggested by Yang, Niu et al. (2011), and Ma et al. (2017): (1) dynamic vegetation (Dickinson et al., 1998; Yang & Niu, 2003), (2) Ball-Berry type canopy stomatal resistance scheme (Ball et al., 1987), (3) Canadian Land Surface Scheme (CLASS) ground snow surface albedo (Verseghy, 1991), (4) Jordan's scheme (Jordan, 1991) for partitioning precipitation into rainfall and snowfall, (5) Noah-type lower boundary condition of soil temperature, and (6) semiimplicit snow/soil temperature time scheme. Table 1 lists the alternative parameterization options for other six processes to generate 288 scheme combinations for subsequent analyses. A brief description of the differences among these options follows:

1. Soil moisture factor β for stomatal resistance (BTR) has three schemes. The notable difference is that the CLM type (Oleson et al., 2004) shows a sharper and narrower range of variation with soil moisture than the Noah type (Chen & Dudhia, 2001) does, and the SSiB type (Xue et al., 1991) is even steeper than the CLM type.
2. Runoff and groundwater (RUN) have four schemes. The SIMGM and SIMTOP are based on TOPMODEL (i.e., TOPography-based hydrological MODEL) to parameterize both surface and subsurface runoff as functions of water table depth, with a simple groundwater model for the former (Niu et al., 2007) and an equilibrium water table for the latter (Niu et al., 2005). The Schaake96 and BATS use free drainage concept for subsurface runoff, with an infiltration-excess-based surface runoff model for the former (Schaake et al., 1996) and a fourth-power function of the top 2-m soil wetness for the latter (Yang & Dickinson, 1996).
3. Surface heat exchange coefficient (SFC) has two schemes. The M-O treats the effect of the roughness lengths with the zero-displacement height d_0 (Brutsaert, 1982) while Chen97 accounts for the difference between roughness lengths for heat and momentum (Chen et al., 1997).
4. Frozen soil permeability (INF) has two schemes in the parameterization of soil hydraulic properties. NY06 defines the soil permeability using soil moisture (Niu & Yang, 2006) while Koren99 uses only the liquid water volume (Koren et al., 1999).
5. Supercooled liquid water (FRZ) in frozen soil has two schemes for the freezing-point depression equation. NY06 takes a more general form (Niu & Yang, 2006), while Koren99 takes a variant form with an extra term that accounts for the increased interface between soil particles and liquid water (Koren et al., 1999).
6. Radiation transfer (RAD) has three schemes to treat the canopy gap with regard to subgrid distributions of vegetation. Scheme 1 parameterizes gap probability as a function of solar zenith angle and the 3-D structure of the vegetation canopy, scheme 2 assumes the gap probability to be zero, and scheme 3 sets the gap probability to one minus the green vegetation fraction (Niu & Yang, 2004).

2.2. Model Input and Setup

The static input data for Noah-MP were derived from various sources: (1) global multiresolution terrain elevation data 2010 (GMTED2010) 30-s topography height, (2) United States Geological Survey (USGS) 30-s land use, (3) Food and Agriculture Organization (FAO) 30-s soil texture, and (4) National Environmental Satellite, Data, and Information Service (NESDIS) 0.144-degree monthly 5-year climatological albedo and green vegetation fraction. We remapped these data onto the model grid using the Weather Research and Forecasting (WRF) model Preprocessing System (WPS) version 3.9.1 (Wang et al., 2017). The soil column was divided into four layers with thicknesses of 0.1, 0.3, 0.6, and 1.0 m from the top to the bottom.

We used the 3-hourly, $0.1° \times 0.1°$ China Meteorological Forcing Dataset (CMFD) developed at the Institute of Tibetan Plateau Research, Chinese Academy of Sciences (He & Yang, 2011; Yang et al., 2010) to drive the Noah-MP. This data set includes the forcing fields of surface pressure, 2-m air temperature and specific humidity, 10-m wind speed, downward shortwave radiation, downward longwave radiation, and

Table 1
Noah-MP Parameterization Schemes Investigated in This Study

Physical process	Scheme
BTR: Soil moisture factor for stomatal resistance	1. Noah: Function of soil moisture, as in Noah (Chen & Dudhia, 2001) 2. CLM: Matric potential related, as in CLM (Oleson et al., 2004) 3. SSiB: Matric potential related, as in SSiB (Xue et al., 1991)
RUN: Runoff and groundwater	1. SIMGM: TOPMODEL-based runoff with the simple groundwater (Niu et al., 2007) 2. SIMTOP: TOPMODEL-based runoff with an equilibrium water table (Niu et al., 2005) 3. Schaake96: Infiltration-excess-based surface runoff with free drainage (Schaake et al., 1996) 4. BATS: BATS runoff with free drainage (Yang & Dickinson, 1996)
SFC: Surface exchange coefficient for heat	1. M-O: Monin-Obukhov similarity theory (Brutsaert, 1982) 2. Chen97: Neglecting the zero displacement height (Chen et al., 1997)
INF: Frozen soil permeability	1. NY06: Function of soil moisture (Niu & Yang, 2006) 2. Koren99: Function of soil liquid water (Koren et al., 1999)
FRZ: Supercooled liquid water	1. NY06: Generalized freezing-point depression (Niu & Yang, 2006) 2. Koren99: Variant freezing-point depression (Koren et al., 1999)
RAD: Radiation transfer	1. gap = *f*(3D, cosz): Canopy gaps from 3-D structure and solar zenith angle (Niu & Yang, 2004) 2. gap=0: No canopy gap (Niu & Yang, 2004) 3. gap = 1−fveg: Gaps from vegetated fraction (Niu & Yang, 2004)

Note. CLM = Community Land Model; SSiB = Simplified Simple Biosphere Model; BATS = Biosphere-Atmosphere Transfer Scheme.

precipitation for the period 1979–2013. CMFD has been widely employed for land surface modeling over China (Chen et al., 2011; Leng et al., 2015; Liu & Xie, 2013; Xue et al., 2013; Zheng et al., 2017).

Our study area was over the entire China land, centered at 35.0°N and 105.0°E with 700 latitudinal × 400 longitudinal grids at an equal 0.1° spacing. Figure 1 illustrates the distributions of elevation, vegetation, and soil conditions, as well as the aridity index, which is defined as the ratio of mean annual precipitation to potential evapotranspiration. Table 2 summarizes the distinct characteristics of the 10 hydrologic regions of China.

To minimize the impact of uncertain initial condition on model performance in an efficient way, a two-stage spin-up was performed to provide different initial conditions for each of the 288 ensemble experiments. First, the 30-year forcing data from 1 January 1981 to 1 January 2011 was cycled 3 times to drive Noah-MP with the physical options as the EXP6 in Yang, Niu et al. (2011), and then the 20-year forcing from 1 January 1981 to 1 January 2001 was applied in each of the 288 experiments to get equilibrium states. By using the final states of these spin-up experiments as initial conditions, 288 ensemble experiments were then integrated, respectively, for another 10-year period from 1 January 2001 to 31 December 2010, results of which were aggregated to monthly for subsequent analyses. All simulations for the spin-up and ensemble experiments were run at a 1-hr time interval.

2.3. Evaluation Data and Metric

We used surface sensible heat (SH) and latent heat (LH) monthly data from the FLUXNET model tree ensemble (MTE) products by Max Planck Institute for Biogeochemistry, available on the global 0.5° grid for the period 1982–2011 (Jung et al., 2010, 2011). These products have been widely employed to evaluate land-atmosphere energy exchanges for LSMs (Bonan et al., 2011; Ma et al., 2017; Xia et al., 2016a; Xia et al., 2016b) and climate models (Anav et al., 2013; Lawrence et al., 2012).

We also used terrestrial water storage anomaly (TWSA) monthly data from the Gravity Recovery and Climate Experiment (GRACE) satellite retrieval products, available on the global 1° grid for the period 2002–2017 (Landerer & Swenson, 2012; Tapley et al., 2004). There are three products from different processing centers, and we adopted their arithmetic mean to reduce measurement errors (Sakumura et al., 2014). This error reduction, however, comes with some signal loss, and hence we applied a gain factor to minimize the loss (Landerer & Swenson, 2012). GRACE data have been widely employed to understand the water cycle for large-scale hydrologic modeling (see a review by Güntner, 2008).

We interpolated the observational products onto our model grid by bilinear interpolation. They were used as the reference for model evaluation. In particular, we adopted Taylor skill score (Taylor, 2001) to quantify the differences between model simulations and observations:

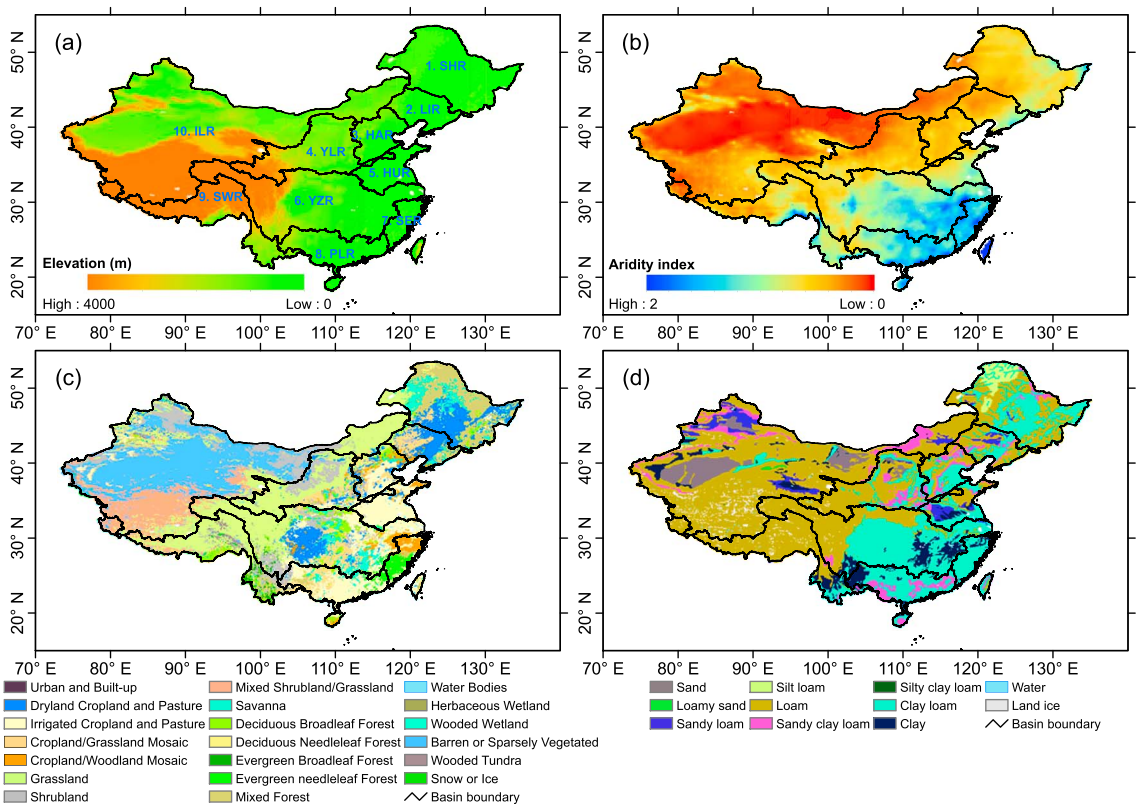

Figure 1. The distributions of (a) elevation, (b) aridity index, (c) vegetation, and (d) soil conditions over 10 hydrologic regions of China.

$$T = \frac{2(1+R)}{\left(\hat{\sigma} + 1/\hat{\sigma}\right)^2} \tag{1}$$

where R is the correlation coefficient and $\hat{\sigma}$ is the ratio of simulated to observed standard deviation. The Taylor skill score ranges from zero to unity, and a higher score indicates a more accurate simulation.

For simultaneous optimization of SH, LH, and TWSA, we aggregated Taylor skill scores into a weighted mean as

$$\overline{T} = \sum_{k=1}^{3} w_k T_k \tag{2}$$

To improve the overall model performance, we assigned larger weights to the variables that were poorly simulated or with lower skill scores

$$w_k = \frac{1 - T_k^{\max}}{\sum_{k=1}^{3} \left(1 - T_k^{\max}\right)} \tag{3}$$

where $T_k^{\max}$ is the maximum Taylor skill score of all ensemble experiments for the kth evaluation variable and $\sum_{k=1}^{3} w_k = 1$.

2.4. Analysis Approach

To objectively examine the differences among the large ensemble of model physical parameterization configurations, a sensitivity score is required. Assume that there are m physical processes (i.e., experimental factors, here $m = 6$) with each process having a different number of parameterization schemes (i.e., treatments). The mean objective function value (i.e., treatment mean) of the jth scheme ($j = 1, 2, \cdots$) of the

Table 2
Properties of the 10 Selected Hydrologic Regions Over China

No.	Hydrologic region	Area (km^2)	Precipitation (mm/year)	Temperature (°C)	Soil texture	Vegetation type
1	Songhua River basin (SHR)	1,105,048.73	517.6	1.4	44.2% loam; 40.6% clay loam; 11.8% silt loam	42.9% mixed forest; 26.3% dryland cropland and pasture; 14.6% savanna; 7.8% grassland
2	Liao River basin (LIR)	344,875.40	542.9	6.9	54.7% loam; 26.6% clay loam; 10.0% sandy clay loam	40.4% dryland cropland and pasture; 27.1% cropland/grassland mosaic; 12.5% mixed forest; 6.4% cropland/woodland mosaic
3	Hai River basin (HAR)	331,541.62	518.4	10.1	42.6% clay loam; 36.6% loam; 20.5% sandy clay loam	31.8% irrigated cropland and pasture; 19.4% cropland/grassland mosaic; 14.6% dryland cropland and pasture; 11.7% cropland/woodland mosaic
4	Yellow River basin (YLR)	803,410.71	468.2	6.6	66.8% loam; 18.8% clay loam; 8.9% sandy clay loam	48.4% grassland; 20.5% shrubland; 11.4% cropland/grassland mosaic; 6.5% irrigated cropland and pasture
5	Huai River basin (HUR)	322,521.09	929.7	14.7	41.6% clay loam; 27.3% loam; 15.3% sandy loam; 9.0% clay	81.8% irrigated cropland and pasture; 6.3% cropland/woodland mosaic; 2.8% dryland cropland and pasture
6	Yangtze River basin (YZR)	1,671,677.13	1,056.1	11.4	50.6% clay loam; 36.5% loam; 10.6% clay	24.4% grassland; 18.4% irrigated cropland and pasture; 11.7% dryland cropland and pasture; 10.6% shrubland; 10.5% cropland/grassland mosaic
7	Southeast River basin (SER)	219,904.27	1,921.9	17.5	75.8% clay loam; 15.1% clay; 4.2% sandy clay loam	33.8% evergreen needleleaf forest; 26.4% cropland/woodland mosaic; 24.1% irrigated cropland and pasture; 10.5% mixed forest
8	Pearl River basin (PLR)	513,435.24	1,537.4	19.8	61.6% clay loam; 23.0% sandy clay loam; 13.9% clay	55.1% irrigated cropland and pasture; 16.5% cropland/grassland mosaic; 5.8% shrubland; 5.3% savanna; 5.2% evergreen needleleaf forest
9	Southwest River basin (SWR)	776,692.78	781.2	4.7	76.7% loam; 9% clay loam; 8.3% clay; 4.6% land ice	46.2% grassland; 12.4% mixed shrubland/grassland; 9.0% shrubland; 8.0% mixed forest; 6.6% deciduous broadleaf forest
10	Inland River basin (ILR)	3,531,123.02	217.8	3.2	67.3% loam; 11.2% sand; 6.6% sandy clay loam; 5.4% sandy loam	41.2% barren or sparsely vegetated; 19.6% grassland; 18.0% mixed shrubland/grassland; 15.4% shrubland

ith process ($i = 1, 2, \cdots, m$) can be represented by $\overline{Y}_j^{(i)}$. We defined an index to quantify the sensitivity of these physical processes as

$$S_i = \frac{\Delta\overline{Y}^{(i)}}{\max\left\{\Delta\overline{Y}^{(1)}, \Delta\overline{Y}^{(2)}, \cdots, \Delta\overline{Y}^{(m)}\right\}} \tag{4}$$

where $\Delta\overline{Y}^{(i)} = \overline{Y}_{\max}^{(i)} - \overline{Y}_{\min}^{(i)}$ is the difference between the largest and the smallest treatment means of the ith process.

To test differences in all possible pairs of means, multiple comparison procedures are often used (Bretz et al., 2010). Suppose that a physical process A has a parameterization schemes (e.g., the physical process RUN has four parameterization schemes) and $\overline{Y}_1, \overline{Y}_2, \cdots, \overline{Y}_a$ are the mean objective function values (i.e., treatment means) of corresponding schemes (i.e., treatments). We adopted the test procedure of Tukey (1949), which has been highly recommended for its computational simplicity and statistical features (Benjamini, 2010). It makes use of the distribution of the studentized range statistic

$$q = \frac{\overline{Y}_{\max} - \overline{Y}_{\min}}{\sqrt{\frac{MS_E}{n}}} = \frac{\overline{Y}_{\max} - \overline{Y}_{\min}}{\sqrt{\frac{SS_E}{n(N-a)}}} \tag{5}$$

where $\overline{Y}_{\max}$ is the largest treatment mean, $\overline{Y}_{\min}$ is the smallest treatment mean, n is the sample size of each treatment, N is the total sample size of all treatments, MS_E is the mean square error, $SS_E = \sum_{p=1}^{a}\sum_{q=1}^{n}$

$\left(Y_{p,q}-\overline{Y}_p\right)^2$ is the sum of the square error, and $N-a$ is the degree of freedom associated with MS_E. Tukey's test declares that the two means are significantly different at the significance level α $(0 \le \alpha \le 1)$ if

$$\left|\overline{Y}_k-\overline{Y}_l\right|>q_\alpha(a,N-a)\sqrt{\frac{MS_E}{n}} \tag{6}$$

where $q_\alpha(a,N-a)$ can be obtained from the critical value table of the studentized range distribution.

Tukey's test is based on the assumptions that the samples are mutually independent and normally distributed with equal variances. We checked the normality by Kolmogorov-Smirnov test (Massey, 1951) and the homogeneity by Levene's test (Levene, 1960). When these two assumptions were violated, we applied the transformation of Box and Cox (1964) to normalize the data with constant variance. We set α to 0.05 and therefore distinguished the categories of the parameterization schemes by Tukey's test at 95% confidence level.

3. Results and Discussion

3.1. Evaluation of the Ensemble Experiments

We evaluated SH, LH, and TWSA from the 288 experiments against observations. The Noah-MP-simulated terrestrial water storage consists of canopy-intercepted water, snow water equivalent, total soil water depth, and groundwater storage (Ma et al., 2017). These components shall be evaluated separately to identify the specific sources of model deficiencies (Xia et al., 2017). However, we focused on comparing total TWSA due to the lack of high-quality observations for individual components in China. Following Xia et al. (2016b), Xia et al. (2017), and Ma et al. (2017), we excluded the water storage in lakes and rivers, which are not modeled in Noah-MP. We then calculated monthly TWSA relative to the same baseline time-average (i.e., 2004–2009) as GRACE-derived TWSA for the period 2003–2010. Finally, we used the same regional mask at the model spatial resolution to derive regional averages of the observed and simulated monthly SH, LH, and TWSA over all the hydrologic regions.

Figure 2 shows the multiyear mean annual cycle of spatially averaged SH, LH, and TWSA for the ensemble experiments and observations over 10 hydrologic regions. The boxplots present the uncertainty ranges by five statistics (i.e., the minimum, first quartile, median, third quartile, and the maximum) of the ensemble experiments. The ensemble experiments basically captured the annual cycle of observations with relatively wide uncertainty ranges in the warm season. This indicates that the performance differences of the selected schemes are more distinct in the warm season.

The seasonal variability of SH was largest over region 10, followed by regions 1–4, medium over regions 5 and 9, and smallest over regions 6–8. This decreasing trend from cold, arid to hot, humid regions is consistent with the finding of Zhou and Huang (2014), and it is attributable to the impacts of the land-air temperature difference and surface wind speed. Most experiments underestimated SH in all seasons over all hydrologic regions. The most obvious underestimation occurred over the most humid regions 6–8 in all seasons and over other regions in the warm rainy season. Pilotto et al. (2015) also found that the underestimation of SH is more significant in the rainy season than in the dry season due to the larger cold bias of soil temperature. Ma et al. (2017) found that Noah-MP underestimates SH throughout the year over most arid regions and greatly underestimates SH in the early growing season over the humid regions. They demonstrated that the underestimation can be alleviated by using a leaf area index climatology prescribed for each land use type.

The seasonal variability of LH was largest over regions 1–3 and 5–8, medium over regions 4 and 9, and smallest over region 10. This decreasing trend from humid to arid regions shows strong latitudinal gradients corresponding to climate patterns in China and is in line with the findings of Yao et al. (2013) and Ma et al. (2019). The ensemble range of LH generally contained observations in most seasons but underestimated LH in winter over all hydrologic regions. This is likely related to the model's failure to reproduce direct root water uptake from the saturated zone, which is a dominant process controlling evapotranspiration of the phreatophytes (Wang et al., 2018; Zheng et al., 2015). In addition, LH underestimation can also be detected in summer over regions 1–3 and 5 and in spring and early summer over regions 7, 8, and 10.

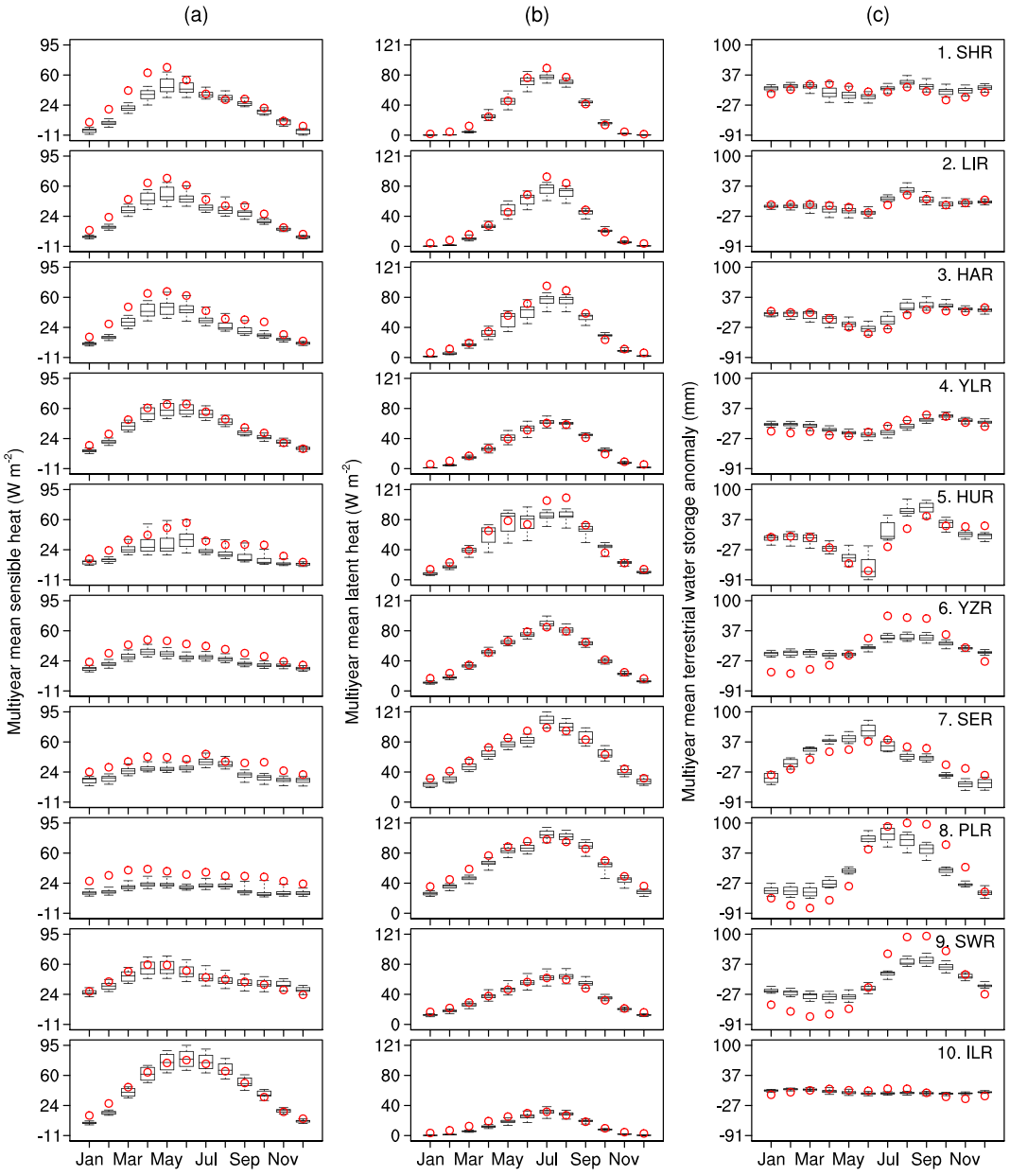

Figure 2. Multiyear mean annual cycle of spatially averaged (a) sensible heat, (b) latent heat, and (c) terrestrial water storage anomaly over 10 hydrologic regions, where box plots are results simulated by 288 ensemble experiments and red circles are observations.

It could be attributed to the dry biases of soil liquid water content in the topsoil (Gao et al., 2015). By contrast, LH was overestimated in summer over regions 6–9. This LH overestimation compensates the concurrent large SH underestimation.

The seasonal variability of TWSA was largest over regions 8 and 9, followed by regions 5–7, medium over regions 1–4, and smallest over region 10. The regional differences agree well with the spatial patterns of precipitation, which are mainly determined by the East Asian monsoon system and the topographic effects (Zhai et al., 2005). Additionally, regional characteristics such as topography, soil, and vegetation conditions also influence TWSA by partitioning water storage in soil and river systems. The ensemble experiments overestimated TWSA in the cold season but underestimated it in the warm season over most regions. One exception was for the regions 2, 3, and 5, where TWSA was underestimated in the cold season but overestimated in the warm season. The large uncertainty ranges for overestimation or underestimation of TWSA are mainly

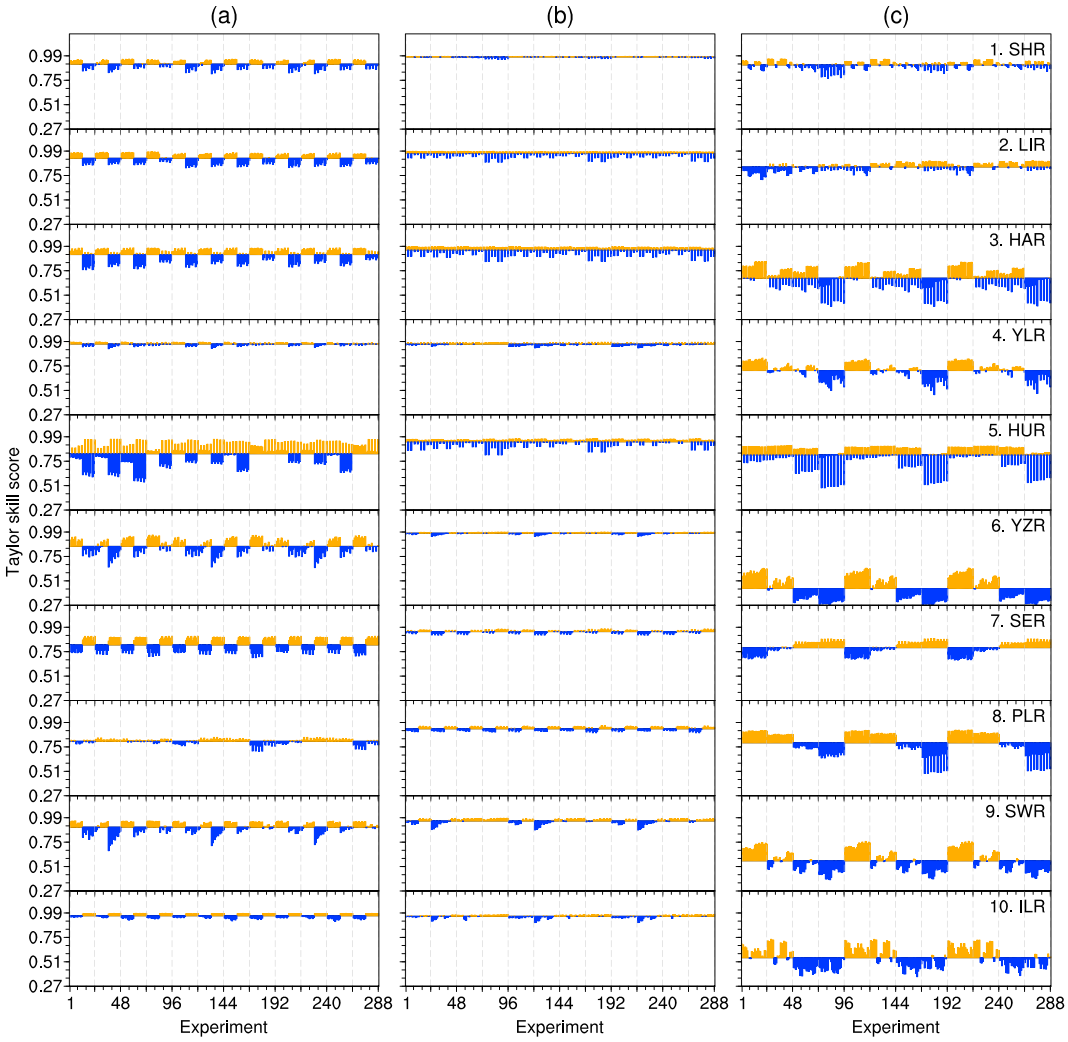

Figure 3. Taylor skill score for (a) sensible heat, (b) latent heat, and (c) terrestrial water storage anomaly over 10 hydrologic regions simulated by 288 ensemble experiments. Orange bars are Taylor skill scores larger than the ensemble mean, while blue bars are those smaller than the ensemble mean.

due to the variability of different scheme combinations in simulating soil moisture, runoff, and groundwater. Besides, snow water equivalent is an important water storage term for the cold regions of China. It not only influences the estimates of terrestrial water storage but also affects total runoff and soil moisture via snowmelt (Xia et al., 2017). The large uncertainty ranges in the cold season for the cold regions may be related to the variability of different scheme combinations for the simulation of snowpack-snowmelt and soil freeze-thaw processes, which are controlled by ground heat flux (Yang, Niu, et al., 2011).

3.2. Uncertainty Analysis of the Ensemble Experiments

Figure 3 compares the performances of the ensemble experiments over 10 hydrologic regions using Taylor skill score. The SH scores ranged between 0.55 and 0.99, with the mean decreasing from 0.97 over regions 4 and 10, 0.90–0.92 over regions 1–3 and 9, to 0.81–0.85 over regions 5–8. The LH scores ranged between 0.81 and 0.99, with the mean of 0.95–0.98 over all regions, except for a reduction to 0.93 over region 8. Thus, the ensemble can well reproduce observed SH and LH variations over all regions. By contrast, TWSA had worse scores, ranging between 0.27 and 0.96, with the mean decreasing from 0.90 over region 1, 0.80–0.84 over regions 2, 5, 7, and 8, 0.55–0.71 over regions 3, 4, 9, and 10, to 0.43 over region 6.

Figure 4 compares Taylor skill scores among different parameterization schemes for each of the six processes. The SH scores exhibited similar behaviors in the SFC process over regions 1–6 and 9, where M-O

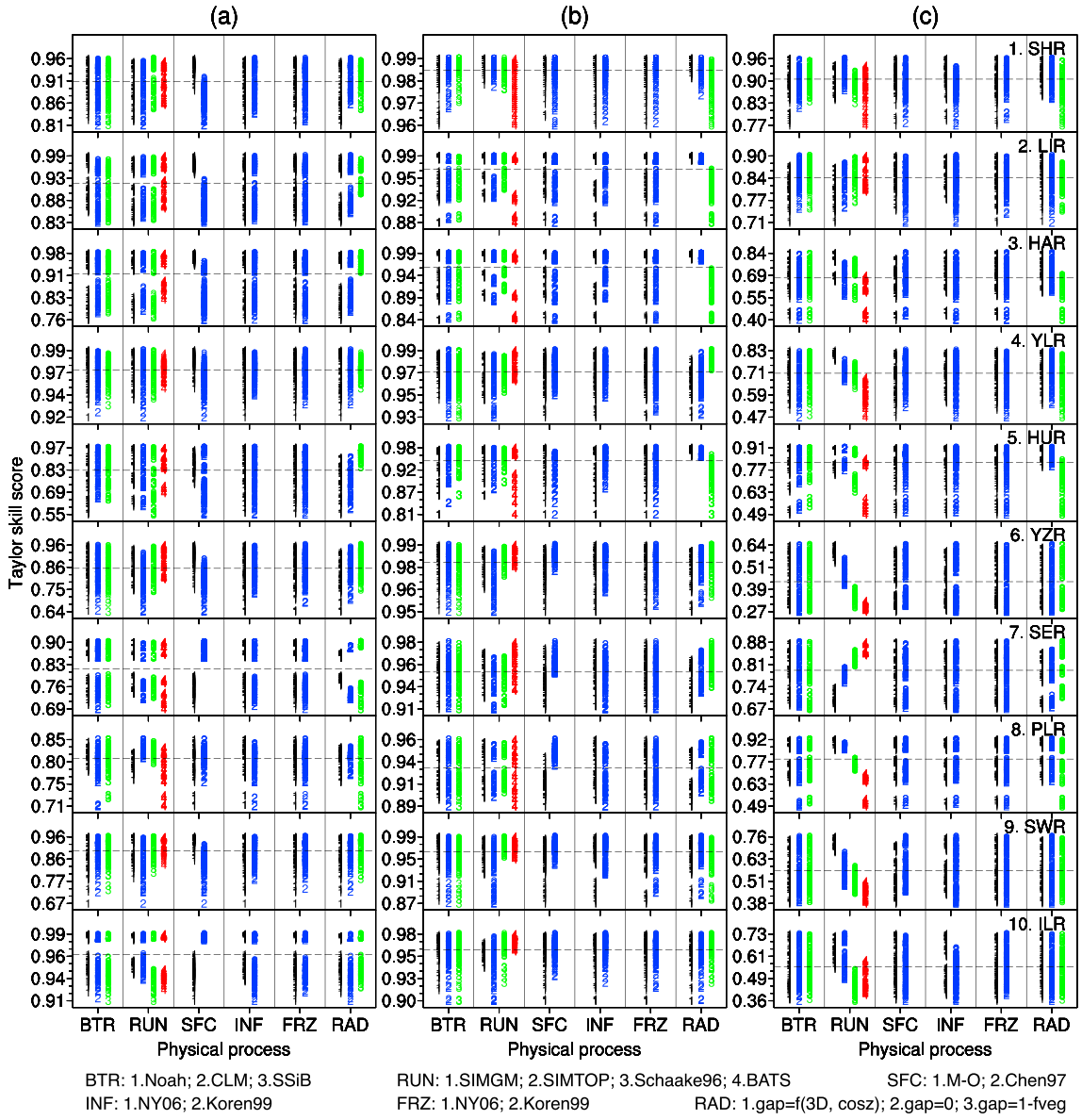

Figure 4. The relationship between parameterization scheme and Taylor skill score for (a) sensible heat, (b) latent heat, and (c) terrestrial water storage anomaly over 10 hydrologic regions simulated by 288 ensemble experiments, where numbers with different colors represent parameterization schemes and dashed lines are the mean Taylor skill scores of all ensemble experiments. BTR = soil moisture factor for stomatal resistance; RUN = runoff-groundwater; SFC = surface heat exchange coefficient; INF = frozen soil permeability; FRZ = supercooled liquid water; RAD = radiation transfer.

scheme performed better than Chen97 scheme. Conversely, Chen97 outperformed M-O over regions 7, 8, and 10. Since M-O using zero-displacement height produces greater SH than Chen97 (Yang, Niu, et al., 2011), it can partly alleviate underestimation, and conversely, Chen97 can reduce overestimation. The SH scores over regions 1–6 also depended on the RAD process, where setting gap to (1−fveg) generally outperformed f(3D, cosz) and zero. This is reasonable since the RAD process controls the radiation transfer through the vegetation canopy with regard to subgrid distributions of vegetation (Niu et al., 2011). The (1−fveg) scheme yields the largest undercanopy solar radiation, while the zero scheme gives the smallest. Moreover, substantial SH score differences existed between the parameterization schemes of the RUN process over regions 1, 4–6, and 8–10, where the best schemes varied a lot with regions. Among the four schemes, SIMTOP produces the wettest soil and greatest evapotranspiration while BATS produces the driest soil and smallest evapotranspiration (Yang, Niu, et al., 2011).

For LH simulation, substantial score differences occurred among the four schemes in the RUN process, where BATS performed the best over regions 4, 6, 7, 9, and 10 but worst over other regions by underestimation. Therefore, the RUN process influences the partitioning of available energy into SH and LH through soil moisture (Xia et al., 2014). The RAD process dominated the model performance over regions 1–5, where setting gap to (1−fveg) outperformed the other two schemes over region 4 but performed worse over the other four regions. Regions with smaller green vegetation fraction have a larger between-canopy gap, exposing more understory vegetation or snow surface to solar radiation and thus producing larger LH by (1−fveg) scheme. Substantial differences were also simulated over regions 4 and 6–10 by the SFC process, where Chen97 outperformed M-O. This occurred because M-O produces smaller LH to balance larger SH by using zero-displacement height (Yang, Niu, et al., 2011). Furthermore, large differences were produced over regions 2–6, 9, and 10 by the INF process, where Koren99 outperformed NY06. The NY06 produces higher permeability in frozen soil and thus more soil water and less surface runoff than Koren99 (Niu & Yang, 2006). Therefore, the INF process influences LH simulation through soil moisture over the regions with frozen soil.

The TWSA score differences were similar over all regions between the schemes in the INF process, where NY06 was generally better than Koren99. The INF process also influenced the water storage, where NY06 reduced TWSA overestimation in the cold season. Substantial score differences were identified over all regions with the RUN process, where BATS performed the best over regions 2 and 7 whereas the worst over other regions. The RUN process directly influences the partitioning of water storage into the soil and river systems (Zheng et al., 2017). Moreover, substantial differences were observed in the SFC process over region 9 where Chen97 outperformed M-O, and in the RAD process over regions 1, 3–5, and 8 where setting gap to (1−fveg) performed worse than the other two schemes. Chen97 produces more LH and thus less soil moisture than M-O, whereas (1−fveg) produces more LH and thus less soil moisture than the other two schemes. Hence, the SFC and RAD processes indirectly affect TWSA through LH.

3.3. Sensitivity Analysis of the Physical Processes

Figure 5 compares the sensitivity scores, based on the average Taylor skill score over all experiments using the same scheme, for SH, LH, and TWSA over 10 hydrologic regions. The sensitivity categories of the physical processes agreed fairly well with the performance differences of the parameterization schemes as identified in Figure 4. However, the sensitivity scores can also identify the insensitive or sensitive processes that are hard to be distinguished by direct comparison of the scheme performances. For instance, the INF process has two schemes that produced significant differences for SH and LH over region 4 (Figure 4) but was insensitive as compared to other processes (Figure 5). This is because only a few experiments using NY06 resulted in worse performance than all others using Koren99, and their performance differences were statistically insignificant. Overall, the number of sensitive processes varied from one to three over different regions, and the FRZ process was insensitive over all regions since its two schemes exhibited nonsignificant performance difference.

For SH simulation, the RUN, SFC, and RAD processes dominated skill score differences, whereas the BTR, INF, and FRZ processes had little influence. SFC was the most sensitive process except for regions 5 and 8, where, respectively, RAD and RUN took over. This agrees with Yang, Niu et al. (2011) in that SFC is the most important factor for SH. The importance of SFC in controlling the total surface energy flux to the atmosphere (i.e., SH and LH) has been shown by many other studies (Chen & Zhang, 2009; Gao et al., 2015; Niu et al., 2011).

For LH simulation, the RUN, SFC, and RAD processes dominated skill score differences over most regions, except that the BTR process also had pronounced influence over regions 1 and 4. On the other hand, the influences of the INF and FRZ processes were negligible. Our result is in general accordance with Hong et al. (2014), who found the sensitivity order for evapotranspiration simulation as RAD > SFC > RUN > BTR > INF > FRZ.

For TWSA simulation, the RUN and RAD processes dominated skill score differences over almost all regions, whereas the SFC and FRZ processes had little influence. Meanwhile, BTR was sensitive over region 2 and INF was sensitive over regions 1 and 2. Our result is similar to Hong et al. (2014), who found the sensitivity order for runoff simulation as RUN > SFC > RAD > INF > BTR > FRZ. The difference is that

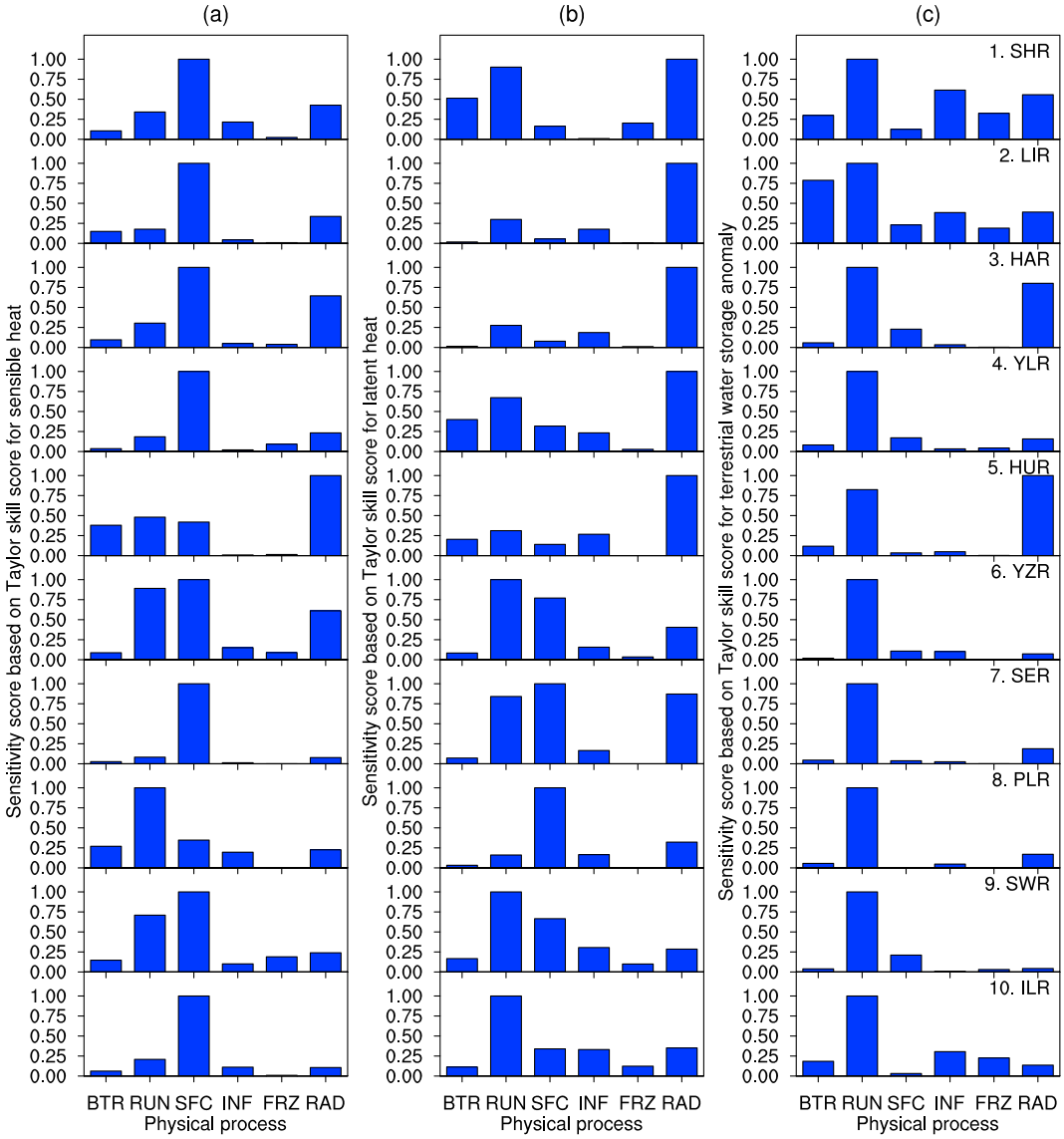

Figure 5. Sensitivity scores of different physical processes for (a) sensible heat, (b) latent heat, and (c) terrestrial water storage anomaly over 10 hydrologic regions simulated by 288 ensemble experiments. BTR = soil moisture factor for stomatal resistance; RUN = runoff-groundwater; SFC = surface heat exchange coefficient; INF = frozen soil permeability; FRZ = supercooled liquid water; RAD = radiation transfer.

we found SFC to be less sensitive than RAD over almost all regions, while INF was more sensitive than RAD over the regions with frozen soil (regions 1, 2, and 10).

3.4. Combinatorial Optimization of the Parameterization Schemes

Based on the scheme categories from multiple comparisons (supporting information Figures S1–S4), we selected the best schemes for optimization of SH, LH, and TWSA independently and simultaneously in each of the 10 hydrologic regions (Table 3). Although the best schemes so selected were generally consistent with those from direct performance comparison (Figure 4), multiple comparisons provide a clearer way of distinguishing parameterization schemes. The best schemes varied with regions of characteristic hydroclimate, soil, and vegetation conditions. Depending on regions, the number of best experiments varied from 2–96 for SH, 2–36 for LH, 4–64 for TWSA, and 2–16 for their simultaneous optimization. The best schemes also varied with variables, especially in the most sensitive processes. For instance, the best RUN schemes were Schaake96 and BATS for SH but SIMGM and SIMTOP for LH over region 1. Thus, optimizing a

Table 3
Best Schemes Based on the Multiple Comparisons of Different Physical Processes for Different Regions and Variables

Optimization variable	Hydrologic region	BTR	RUN	SFC	INF	FRZ	RAD	Number of best experiments
SH	1. SHR	1	3, 4	1	2	1, 2	2	4
	2. LIR	1	4	1	2	1, 2	3	2
	3. HAR	2, 3	4	1	1, 2	1, 2	3	8
	4. YLR	1, 2, 3	1, 2, 3, 4	1	1, 2	1, 2	2, 3	96
	5. HUR	2, 3	4	1	1, 2	1, 2	3	8
	6. YZR	1, 2	4	1	2	2	3	2
	7. SER	1, 2, 3	1, 2, 3	2	1, 2	1, 2	1, 3	72
	8. PLR	1	2, 3	2	1	1, 2	2, 3	8
	9. SWR	2, 3	4	1	2	2	2, 3	4
	10. ILR	1, 2	1, 2	2	1	1, 2	3	8
LH	1. SHR	2, 3	1, 2	1	1, 2	2	1	8
	2. LIR	1, 2, 3	1, 2, 3	1	2	1, 2	1, 2	36
	3. HAR	1, 2, 3	1, 3	1	2	1, 2	1, 2	24
	4. YLR	1	4	2	2	1, 2	3	2
	5. HUR	2, 3	1, 2, 3	1	2	1, 2	1, 2	24
	6. YZR	1, 2, 3	4	2	2	1, 2	3	6
	7. SER	1, 2, 3	4	2	1	1, 2	3	6
	8. PLR	1, 2, 3	1, 2, 3	2	1	1, 2	2	18
	9. SWR	1	3, 4	2	2	2	2	2
	10. ILR	1, 2, 3	4	2	2	2	3	3
TWSA	1. SHR	2, 3	2	1	1	2	1, 2	4
	2. LIR	2, 3	3, 4	1	1	2	1, 2	8
	3. HAR	1, 3	1	2	1	1, 2	1, 2	8
	4. YLR	1, 3	1	2	1	1	1, 2	4
	5. HUR	2, 3	1, 2	1, 2	1, 2	1, 2	1, 2	64
	6. YZR	2, 3	1	2	2	1, 2	3	4
	7. SER	1, 2	4	2	1	1, 2	3	4
	8. PLR	1, 3	1	1, 2	2	1, 2	1, 2	16
	9. SWR	2, 3	1	2	1, 2	2	2, 3	8
	10. ILR	2, 3	1, 2	1, 2	1	2	1	8
SH, LH, and TWSA	1. SHR	2, 3	2	1	1	2	2	2
	2. LIR	2, 3	3, 4	1	1	2	1, 2	8
	3. HAR	1, 3	1	2	1	1, 2	2	4
	4. YLR	1, 3	1	2	1	1	1, 2	4
	5. HUR	2, 3	1, 2	1	2	1, 2	1, 2	16
	6. YZR	1, 2, 3	1	2	2	1, 2	3	6
	7. SER	1, 2	4	2	1	1, 2	3	4
	8. PLR	1, 3	2	2	1, 2	1, 2	2	8
	9. SWR	2, 3	1	2	1, 2	2	2, 3	8
	10. ILR	2, 3	1, 2	1, 2	1	2	1	8

Note. SH = sensible heat; LH = latent heat; TWSA = terrestrial water storage anomaly; BTR = soil moisture factor for stomatal resistance; RUN = runoff-groundwater; SFC = surface heat exchange coefficient; INF = frozen soil permeability; FRZ = supercooled liquid water; RAD = radiation transfer.

single variable can sometimes degrade the skill in simulating other variables (Hong et al., 2014). Overall, the best schemes for the simultaneous optimization of SH, LH, and TWSA were more consistent with those for the independent optimization of TWSA because a larger weight was assigned to TWSA than SH and LH. However, the best schemes for SH and LH were also so identified for simultaneous optimization of the three variables even though they were not optimal for TWSA, that is, the Noah scheme in the BTR process over regions 6, the SIMTOP scheme in the RUN process over region 8, and the NY06 scheme in the INF process over regions 8. On the other hand, the worst schemes for SH and/or LH were not identified as the best schemes for simultaneous optimization even if they were optimal for TWSA, that is, the Chen97 scheme in the SFC process over region 5, the M-O scheme in the SFC process over region 8, the NY06 scheme in the INF process over region 5, and the f(3D, cosz) scheme in the RAD process over regions 1, 3, and 8. The outcome can be explained by the fact that the performance differences between those and other schemes of the same process were relatively less obvious for TWSA than for SH and/or LH.

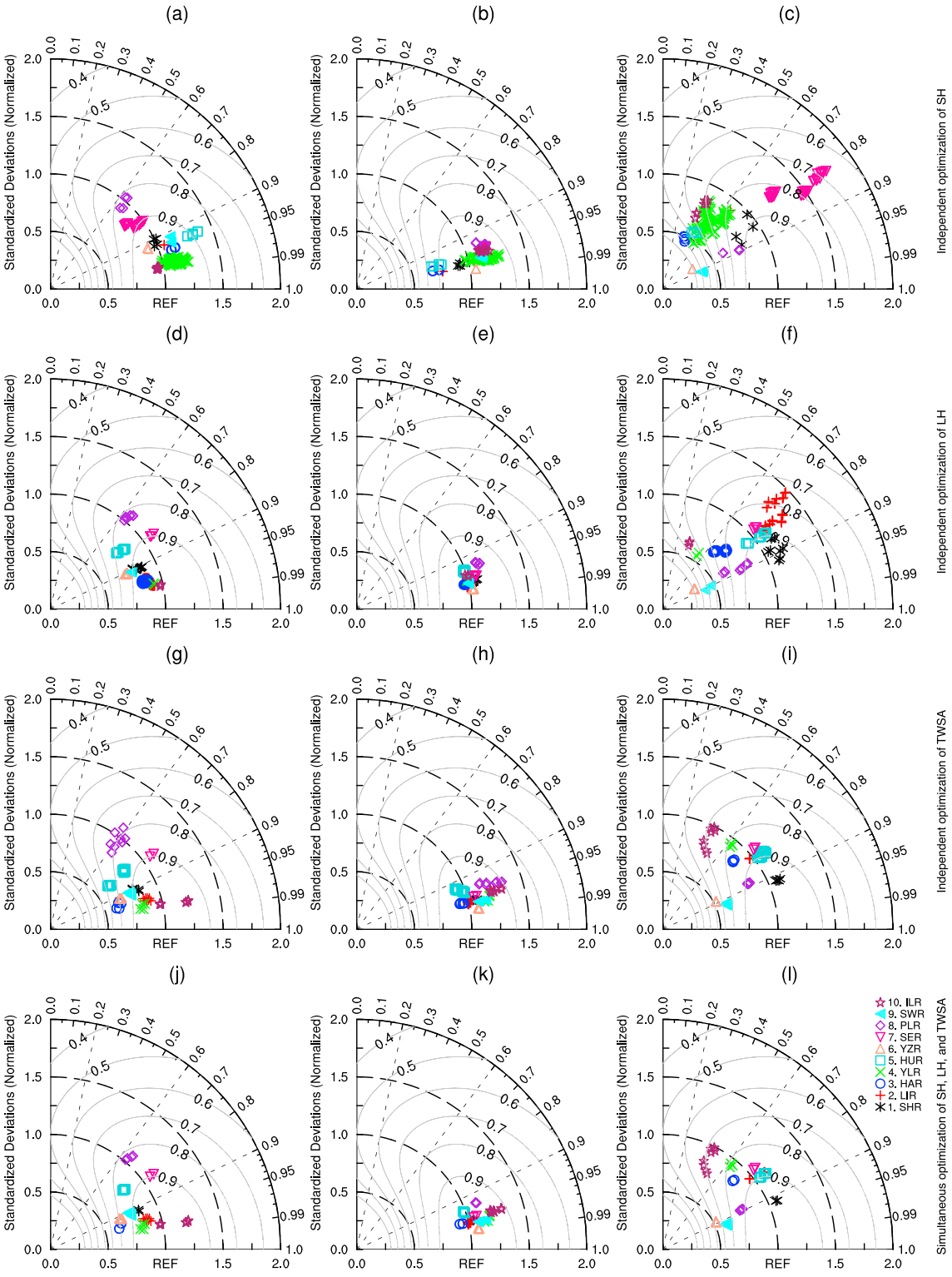

Figure 6. Taylor diagrams of the best experiments for the simulation of sensible heat (first column, a, d, g, and j), latent heat (second column, b, e, h, and k), and terrestrial water storage anomaly (third column, c, f, i, and l) based on independent optimization of sensible heat (first row, a--c), latent heat (second row, d--f), and terrestrial water storage anomaly (third row, g--i) and simultaneous optimization of all three variables (fourth row, j--l). Gray solid lines represent the Taylor skill score contours. SH = sensible heat; LH = latent heat; TWSA = terrestrial water storage anomaly.

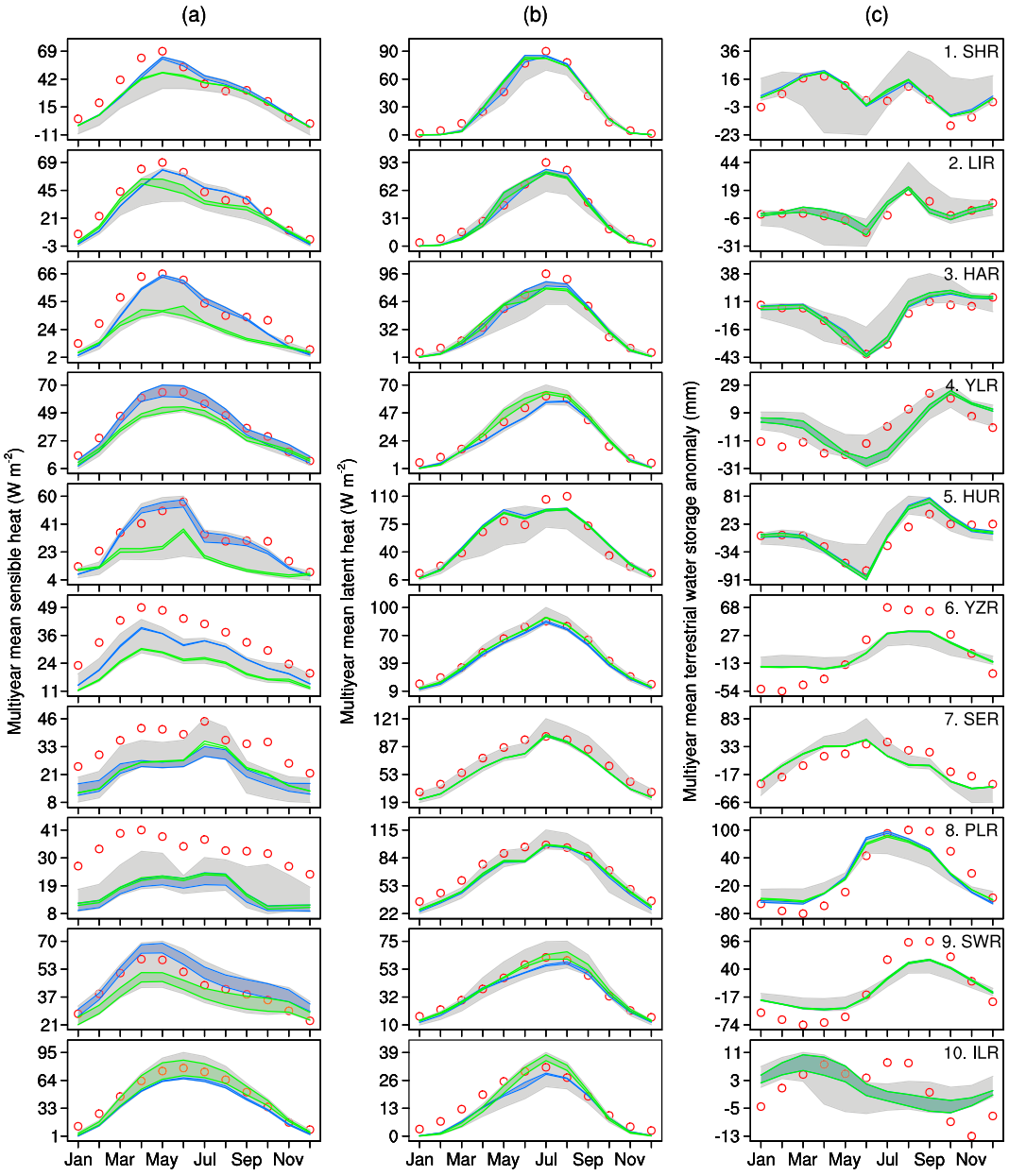

Figure 7. Multiyear mean annual cycle of spatially averaged (a) sensible heat, (b) latent heat, and (c) terrestrial water storage anomaly over 10 hydrologic regions, where gray bands are results simulated by 288 ensemble experiments, blue bands are results simulated by the best experiments of independent optimization, green bands are results simulated by the best experiments of simultaneous optimization, and red circles are observations.

Figure 6 presents the Taylor diagrams of the simulations by the best parameterization scheme combinations (i.e., best experiments) based on independent SH, LH, and TWSA optimization and their simultaneous optimization as shown in Table 3. The best experiments always had larger Taylor skill scores for LH than SH and TWSA, regardless which optimization was used. The independent optimization of one variable reduced the scores of other variables. For example, the SH optimization (Figure 6a) reduced the scores for LH (Figure 6b) and TWSA (Figure 6c) when comparing with LH optimization (Figure 6e) and TWSA optimization (Figure 6i), respectively. The simultaneous optimization allowed tradeoffs among all variables and produced overall more reasonable results (Figures 6j–6l), although it inevitably reduced the scores over some regions for some variables. The score reduction was much obvious for SH over regions 3, 6, and 9 (comparing Figure 6a with Figure 6j), for LH over regions 4, 9, and 10 (comparing Figure 6e with Figure 6k), and for TWSA over region 8 (comparing Figure 6i with Figure 6l).

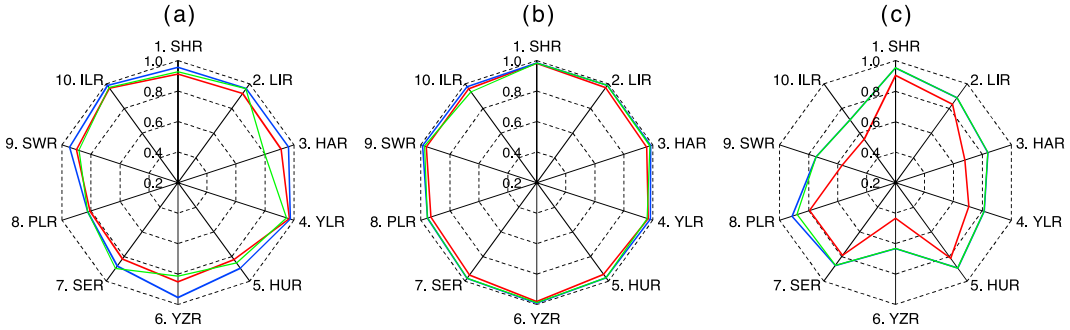

Figure 8. Mean Taylor skill scores for (a) sensible heat, (b) latent heat, and (c) terrestrial water storage anomaly over 10 hydrologic regions, where red lines are results simulated by 288 ensemble experiments, blue lines are results simulated by the best experiments of independent optimization, and green lines are results simulated by the best experiments of simultaneous optimization.

3.5. Comparison Between the Ensemble and Best Experiments

Figure 7 compares multiyear mean annual cycles of SH, LH, and TWSA averaged over each of the 10 hydrologic regions from all 288 experiments (i.e., ensemble experiments) and those using the best parameterization scheme combinations (i.e., best experiments) based on independent and simultaneous optimizations. The uncertainty ranges of the best experiments were much narrower and closer to observations than those of the ensemble experiments. The result based on the simultaneous optimization agreed better with that from TWSA-independent optimization giving its larger weight than SH and LH. The optimized Noah-MP can better capture the overall water and energy cycles, including both timing and magnitude. Nonetheless, there were still large SH underestimations over the most humid regions 6–8, which are related to cold soil temperature biases (Pilotto et al., 2015) and so unable to be reduced through the optimization of the existing parameterizations. Meanwhile, the relatively large LH bias over the highly irrigated cropland region 5 is likely due to the use of leaf dynamics in Noah-MP, which may capture crop growth under rainfed but not irrigated conditions (Cai, Yang, David, et al., 2014).

Figure 8 compares the mean Taylor skill scores of SH, LH, and TWSA for ensemble experiments and best experiments based on independent and simultaneous optimizations. The score gain was the largest from TWSA optimization, about 0.2 (46%) over region 6, 0.16–0.18 (24–32%) over regions 3, 9, and 10, 0.07–0.12 (10–15%) over regions 4, 5, 7, and 8, and 0.05 (6%) over regions 1 and 2. The gain from SH optimization was about 0.06–0.1 (7–12%) over regions 5–7, 0.05 (5%) over regions 1–3 and 9, and less than 0.02 (3%) over other regions. Although the gain from LH optimization was not significant (less than 0.03 or 3%), its score was already larger than SH and TWSA. The score from simultaneous optimization agreed well with that from TWSA optimization over all regions. Because of the tradeoff, the simultaneous optimization reduced the SH score respectively by 0.16 (18%), 0.14 (17%), and 0.07 (7%) over regions 3, 6, and 9, reduced the LH score respectively by 0.02 (2%), 0.02 (1%), and 0.04 (4%) over regions 4, 9, and 10, and reduced the TWSA score by 0.03 (4%) over region 8.

The low discrepancy of SH and LH from observations implies that the surface energy budget was well reproduced by the best parameterization combinations. The relatively large TWSA discrepancy is attributable to model biases in soil moisture, groundwater, and snow water equivalent. These biases are likely related to the model assumption of evenly distributed soil texture (Wang et al., 2018), which could be improved by considering vertical heterogeneity in soil layers (Gao et al., 2015) and more so by incorporating three-dimensional subgrid variability (Choi et al., 2007). The Noah-MP's disagreement with GRACE observations may be partly due to its lack of lake water storage (Ma et al., 2017). In addition, Noah-MP simulates only the natural part of TWSA without considering anthropogenic impacts such as irrigation and reservoir regulation, which were detected by GRACE (Cai, Yang, David, et al., 2014; Ma et al., 2017). Furthermore, GRACE data contain measurement, aliasing, and signal leakage errors that can also lead to the disagreement (Güntner, 2008; Seo et al., 2006). Obviously, the overall performance of any LSM, as evaluated standalone, is highly dependent on the quality of the forcing data, especially precipitation and radiation, which play an essential role in surface flux exchanges through impacts on topsoil temperature and moisture (Gao et al., 2015). The model discrepancies from the observational reference may, therefore, be a result of the uncertainties in the

reference itself as well as the errors in the forcing data, which also account for the nonclosure of the surface water and energy budgets. For example, the FLUXNET MTE products have an underestimated interannual variability because some explanatory variables that used to train the MTE were assumed static over years (Jung et al., 2011). Nonetheless, some parameterization schemes and/or parameters have considerable impacts on model performance and can be tuned to better represent the regional hydroclimatic regimes and watershed properties (Gan et al., 2014, 2018). Our analysis above has further demonstrated how to identify sensitive processes and best parameterizations for a specific LSM.

4. Summary and Conclusions

We conducted sensitivity and uncertainty analyses on the community LSM, Noah-MP, to identify the most responsive processes and the best parameterization combinations for simulating SH, LH, and TWSA over 10 hydrologic regions in China. For this, we designed and conducted an ensemble of 288 parameterization experiments, combining six physical processes (BTR, RUN, SFC, INF, FRZ, and RAD) each with multiple alternative schemes. We used Taylor skill score to quantify the model performance. The ensemble generally captured the observed seasonal variability of surface energy and water budgets. The model simulated LH realistically, with scores ranging from 0.81–0.99 among regions, underestimated SH systematically, with scores between 0.55 and 0.99, and overestimated (underestimated) TWSA in the cold (warm) season over most regions, with scores reduced to 0.27–0.96. Direct comparison of these scores allows us to detect significant performance differences among alternative parameterization schemes in terms of specific processes.

However, performance differences among various schemes across multiple processes were difficult to be distinguished by direct comparison. Sensitivity scores based on the difference between the largest and smallest mean Taylor skill scores were able to identify the relative degrees of sensitivity across all physical processes. We found that the process sensitivities were region-specific because the performances of different parameterization schemes were depended on regional characteristics such as hydroclimate, soil, and vegetation conditions. In particular, the RUN, SFC, and RAD processes were most sensitive for SH and LH over almost all regions. By contrast, the TWSA was dominated by the RUN and RAD processes while largely influenced by the BTR and INF processes over some limited regions. On the other hand, the FRZ process was relatively insensitive over all regions for all variables.

We further used Tukey's test to identify the best parameterization schemes by multiple comparisons of the mean Taylor skill scores for different schemes of the same process. The independent optimization of individual variables improved the mean scores by 0.6–12.1% for SH, 0.1–3.1% for LH, and 5.4–45.5% for TWSA, comparing with those of ensemble experiments. No single scheme combination performed well over all regions or all variables. The best scheme combination from one region may perform poorly in other regions, and improving the skill for one variable may reduce the score for other variables. As such, the simultaneous optimization made tradeoffs among all variables, which generally reduced the skill enhancement from independent optimization, but still improved the overall model performance over the ensemble. The score gains were −12.1–9.5% for SH, −2.6–3.0% for LH, and 5.6–45.5% for TWSA, comparing with those of the ensemble mean.

Although the sensitive physical processes and the best parameterization schemes varied with regions and variables, our systematic analyses allow us to recommend the best Noah-MP configurations for simulating surface water and energy fluxes over China as follows—process (scheme): RUN(4), SFC(1), RAD(3), INF(2), BTR(1), FRZ(2) for SH; RUN(3), SFC(2), RAD(2), INF(2), BTR(1, 2, 3), FRZ(2) for LH; RUN(2), SFC(2), RAD(2), INF(1), BTR(3), FRZ(2) for TWSA; and RUN(2), SFC(2), RAD(2), INF(1), BTR(3), FRZ(2) for their joint.

Caution must be taken for our conclusion and recommendation above. The best configurations may not be good at modeling extremes such as drought and flood, since our emphasis was placed on seasonal rather than interannual variability of the land surface fluxes. We have focused on the model structural uncertainty that arises from alternative parameterization schemes of the selected physical processes, which is only one of the many sources of uncertainty in LSMs (Nearing et al., 2016). We have not tuned specific parameters in these schemes but taken their default values, which are not universal. Some key parameters should be tuned to improve each scheme's performance, especially considering regional surface characteristics (Cuntz et al., 2016; Mendoza et al., 2015). Besides, data uncertainties from the forcing conditions (Xia et al., 2005), soil

textures (Zheng & Yang, 2016), and vegetation variations (Li et al., 2018) can affect model skill scores. Data sets from different sources need to be evaluated and compared to increase our confidence in model simulations. Some significant physics such as subgrid variability (Choi et al., 2007), crop dynamics (Bonan & Doney, 2018), and lake dynamics (MacKay et al., 2009) are still missing in Noah-MP. Future improvement may include these physics to account for lateral flow, rhizosphere effect, dynamic crop growth, surface water storage, and human-induced water use. Nonetheless, our study offers not only a systematic approach for model structural sensitivity and uncertainty analyses but also an optimized set of physics configurations from the existing parameterizations, facilitating further model improvement.

Acknowledgments
This research was supported by the National Key Research and Development Project of China (2017YFC1404000), the National Natural Sciences Foundation of China (41505092 and 41605073), China Meteorological Administration/National Climate Center research subcontract to NUIST (2212031635601), and the National Oceanic and Atmospheric Administration (grant #NA18OAR4590410). The China Meteorological Forcing data set is available at http://dam.itpcas.ac.cn/chs/rs/?q=data#CMFD_0.1, the global observation-based monthly sensible heat and latent heat products are available at https://www.bgc-jena.mpg.de/geodb/projects/Home.php, and the global GRACE-derived terrestrial water storage anomaly products are available at http://grace.jpl.nasa.gov.

References

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., et al. (2013). Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth System Models. *Journal of Climate*, *26*(18), 6801–6843. https://doi.org/10.1175/JCLI-D-12-00417.1

Arsenault, K. R., Nearing, G. S., Wang, S., Yatheendradas, S., & Peters-Lidard, C. D. (2018). Parameter sensitivity of the Noah-MP land surface model with dynamic vegetation. *Journal of Hydrometeorology*, *19*(5), 815–830. https://doi.org/10.1175/jhm-d-17-0205.1

Bai, Y., Wagener, T., & Reed, P. (2009). A top-down framework for watershed model evaluation and selection under uncertainty. *Environmental Modelling & Software*, *24*(8), 901–916. https://doi.org/10.1016/j.envsoft.2008.12.012

Ball, J. T., Woodrow, I. E., & Berry, J. A. (1987). A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. In J. Biggins (Ed.), *Progress in Photosynthesis Research* (pp. 221–234). Dordrecht, Netherlands: Martinus Nijhoff. https://doi.org/10.1007/978-94-017-0519-6_48

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

Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, *52*(6), 708–721. https://doi.org/10.1002/bimj.200900299

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. . L. H., Ménard, C. B., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677–699. https://doi.org/10.5194/gmd-4-677-2011

Bonan, G. B., & Doney, S. C. (2018). Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. *Science*, *359*(6375). https://doi.org/10.1126/science.aam8328

Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., et al. (2011). Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data. *Journal of Geophysical Research*, *116*, G02014. https://doi.org/10.1029/2010JG001593

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *26*(2), 211–252.

Bretz, F., Hothorn, T., & Westfall, P. (2010). *Multiple Comparisons using R*. Boca Raton, FL: Chapman & Hall/CRC.

Brutsaert, W. A. (1982). *Evaporation into the atmosphere: Theory, history and applications*. Dordrecht, Netherlands: Springer Netherlands. https://doi.org/10.1007/978-94-017-1497-6

Cai, X., Yang, Z.-L., David, C. H., Niu, G.-Y., & Rodell, M. (2014). Hydrological evaluation of the Noah-MP land surface model for the Mississippi River basin. *Journal of Geophysical Research: Atmospheres*, *119*, 23–38. https://doi.org/10.1002/2013JD020792

Cai, X., Yang, Z.-L., Fisher, J. B., Zhang, X., Barlage, M., & Chen, F. (2016). Integration of nitrogen dynamics into the Noah-MP Land Surface Model v1.1 for climate and environmental predictions. *Geoscientific Model Development*, *9*(1), 1–15. https://doi.org/10.5194/gmd-9-1-2016

Cai, X., Yang, Z.-L., Xia, Y., Huang, M., Wei, H., Leung, L. R., & Ek, M. B. (2014). 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*, 13,751–13,770. https://doi.org/10.1002/2014JD022113

Chen, F., & Dudhia, J. (2001). Coupling an advanced land surface-hydrology model with the Penn state-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Monthly Weather Review*, *129*(4), 569–585. https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2

Chen, F., Janjić, Z. I., & Mitchell, K. (1997). Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the NCEP mesoscale Eta model. *Boundary-Layer Meteorology*, *85*(3), 391–421. https://doi.org/10.1023/A:1000531001463

Chen, F., Zhang, G., Barlage, M., Zhang, Y., Hicke, J. A., Meddens, A., et al. (2015). An observational and modeling study of impacts of bark beetle-caused tree mortality on surface energy and hydrological cycles. *Journal of Hydrometeorology*, *16*(2), 744–761. https://doi.org/10.1175/JHM-D-14-0059.1

Chen, F., & Zhang, Y. (2009). On the coupling strength between the land surface and the atmosphere: From viewpoint of surface exchange coefficients. *Geophysical Research Letters*, *36*, L10404. https://doi.org/10.1029/2009GL037980

Chen, Y., Yang, K., He, J., Qin, J., Shi, J., Du, J., & He, Q. (2011). Improving land surface temperature modeling for dry land of China. *Journal of Geophysical Research*, *116*, D20104. https://doi.org/10.1029/2011JD015921

Choi, H. I., Kumar, P., & Liang, X.-Z. (2007). Three-dimensional volume-averaged soil moisture transport model with a scalable parameterization of subgrid topographic variability. *Water Resources Research*, *43*, W04414. https://doi.org/10.1029/2006WR005134

Choi, H. I., & Liang, X.-Z. (2010). Improved terrestrial hydrologic representation in mesoscale land surface models. *Journal of Hydrometeorology*, *11*(3), 797–809. https://doi.org/10.1175/2010JHM1221.1

Choi, H. I., Liang, X.-Z., & Kumar, P. (2013). A conjunctive surface-subsurface flow representation for mesoscale land surface models. *Journal of Hydrometeorology*, *14*(5), 1421–1442. https://doi.org/10.1175/JHM-D-12-0168.1

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*, W09301. https://doi.org/10.1029/2010WR009827

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*, 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*, 2515–2542. https://doi.org/10.1002/2015WR017200

Cosgrove, B., Gochis, D. J., Clark, E., Cui, Z., Dugger, A., Feng, X., et al. (2017). Continental-scale operational hydrologic modeling: Version 1.0 of the National Water Model. Paper presented at 97th American Meteorological Society Annual Meeting, Seattle, WA.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., et al. (2016). The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. *Journal of Geophysical Research: Atmospheres*, *121*, 10,676–10,700. https://doi.org/10.1002/2016JD025097

Dai, Y., Zeng, X., Dickinson, R. E., Baker, I., Bonan, G. B., Bosilovich, M. G., et al. (2003). The common land model. *Bulletin of the American Meteorological Society*, *84*(8), 1013–1024. https://doi.org/10.1175/BAMS-84-8-1013

Dickinson, R. E., Henderson-Sellers, A., Kennedy, P. J., & Wilson, M. F. (1986). *Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR community climate model*. Boulder, CO: National Center for Atmospheric Research. https://doi.org/10.5065/D6668B58

Dickinson, R. E., Oleson, K. W., Bonan, G., Hoffman, F., Thornton, P., Vertenstein, M., et al. (2006). The Community Land Model and its climate statistics as a component of the Community Climate System Model. *Journal of Climate*, *19*(11), 2302–2324. https://doi.org/10.1175/JCLI3742.1

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

Duan, Q., Di, Z., Quan, J., Wang, C., Gong, W., Gan, Y., et al. (2017). Automatic model calibration: A new way to improve numerical weather forecasting. *Bulletin of the American Meteorological Society*, *98*(5), 959–970. https://doi.org/10.1175/BAMS-D-15-00104.1

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*, *108*(D22), 8851. https://doi.org/10.1029/2002JD003296

Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., et al. (2014). A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environmental Modelling & Software*, *51*, 269–285. https://doi.org/10.1016/j.envsoft.2013.09.031

Gan, Y., Liang, X.-Z., Duan, Q., Choi, H. I., Dai, Y., & Wu, H. (2015). Stepwise sensitivity analysis from qualitative to quantitative: Application to the terrestrial hydrological modeling of a Conjunctive Surface-Subsurface Process (CSSP) land surface model. *Journal of Advances in Modeling Earth Systems*, *7*, 648–669. https://doi.org/10.1002/2014MS000406

Gan, Y., Liang, X.-Z., Duan, Q., Ye, A., Di, Z., Hong, Y., & Li, J. (2018). A systematic assessment and reduction of parametric uncertainties for a distributed hydrological model. *Journal of Hydrology*, *564*, 697–711. https://doi.org/10.1016/j.jhydrol.2018.07.055

Gao, Y., Li, K., Chen, F., Jiang, Y., & Lu, C. (2015). Assessing and improving Noah-MP land model simulations for the central Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, *120*, 9258–9278. https://doi.org/10.1002/2015JD023404

Gayler, S., Wöhling, T., Grzeschik, M., Ingwersen, J., Wizemann, H. D., Warrach-Sagi, K., et al. (2014). Incorporating dynamic root growth enhances the performance of Noah-MP at two contrasting winter wheat field sites. *Water Resources Research*, *50*, 1337–1356. https://doi.org/10.1002/2013WR014634

Gochis, D. J., Yu, W., & Yates, D. N. (2013). *The WRF-Hydro model technical description and user's guide, version 1.0.* . Boulder, CO: National Center for Atmospheric Research.

Güntner, A. (2008). Improvement of global hydrological models using GRACE data. *Surveys in Geophysics*, *29*(4-5), 375–397. https://doi.org/10.1007/s10712-008-9038-y

He, J., & Yang, K. (2011). China meteorological forcing dataset. *Cold and Arid Regions Science Data Center at Lanzhou*. https://doi.org/10.3972/westdc.002.2014.db

Henderson-Sellers, A., Yang, Z.-L., & Dickinson, R. E. (1993). The Project for intercomparison of land-surface parameterization schemes. *Bulletin of the American Meteorological Society*, *74*(7), 1335–1349. https://doi.org/10.1175/1520-0477(1993)074<1335:TPFIOL>2.0.CO;2

Hong, S., Yu, X., Park, S. K., Choi, Y., & Myoung, B. (2014). Assessing optimal set of implemented physical parameterization schemes in a multi-physics land surface model using genetic algorithm. *Geoscientific Model Development*, *7*(5), 2517–2529. https://doi.org/10.5194/gmd-7-2517-2014

Jordan, R. E. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for SNTHERM.89. Hanover, NH: Cold Region Research and Engineering Laboratory, U.S. Army Corps of Engineers.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*(7318), 951–954. https://doi.org/10.1038/nature09396

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research*, *116*, G00J07. https://doi.org/10.1029/2010JG001566

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*, *104*(D16), 19569–19585. https://doi.org/10.1029/1999JD900232

Krishnakumar, K. (1990). Micro-genetic algorithms for stationary and non-stationary function optimization. Paper presented at 1989 Symposium on Visual Communications, Image Processing, and Intelligent Robotics Systems, SPIE, Philadelphia, PA.

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*, W04531. https://doi.org/10.1029/2011WR011453

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Fletcher, C. G., Lawrence, P. J., Levis, S., et al. (2012). The CCSM4 land simulation, 1850–2005: Assessment of surface climate and new capabilities. *Journal of Climate*, *25*(7), 2240–2260. https://doi.org/10.1175/JCLI-D-11-00103.1

Leng, G., Tang, Q., Huang, M., & Leung, L. R. (2015). A comparative analysis of the impacts of climate change and irrigation on land surface and subsurface hydrology in the North China Plain. *Regional Environmental Change*, *15*(2), 251–263. https://doi.org/10.1007/s10113-014-0640-x

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. Mann (Eds.), *Contributions to probability and statistics: Essay in honor of Harold Hotelling*, (pp. 278–292). Stanford, CA: Stanford University Press.

Li, J., Chen, F., Zhang, G., Barlage, M., Gan, Y., Xin, Y., & Wang, C. (2018). Impacts of land cover and soil texture uncertainty on land model simulations over the central Tibetan Plateau. *Journal of Advances in Modeling Earth Systems*, *10*(9), 2121–2146. https://doi.org/10.1029/2018MS001377

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*, *99*(D7), 14,415–14,428. https://doi.org/10.1029/94JD00483

Liang, X.-Z., Xu, M., Yuan, X., Ling, T., Choi, H. I., Zhang, F., et al. (2012). Regional Climate-Weather Research and Forecasting Model (CWRF). *Bulletin of the American Meteorological Society*, *93*(9), 1363–1387. https://doi.org/10.1175/BAMS-D-11-00180.1

Liu, J. G., & Xie, Z. H. (2013). Improving simulation of soil moisture in China using a multiple meteorological forcing ensemble approach. *Hydrology and Earth System Sciences*, *17*(9), 3355–3369. https://doi.org/10.5194/hess-17-3355-2013

Ma, N., Niu, G.-Y., Xia, Y., Cai, X., Zhang, Y., Ma, Y., & Fang, Y. (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*, 12,245–12,268. https://doi.org/10.1002/2017JD027597

Ma, N., Szilagyi, J., Zhang, Y., & Liu, W. (2019). Complementary-relationship-based modeling of terrestrial evapotranspiration across China during 1982–2012: Validations and spatiotemporal analyses. *Journal of Geophysical Research: Atmospheres*, *124*, 4326–4351. https://doi.org/10.1029/2018JD029850

MacKay, M. D., Neale, P. J., Arp, C. D., de Senerpont Domis, L. N., Fang, X., Gal, G., et al. (2009). Modeling lakes and reservoirs in the climate system. *Limnology and Oceanography*, *54*(6part2), 2315–2329. https://doi.org/10.4319/lo.2009.54.6_part_2.2315

Manabe, S. (1969). Climate and the ocean circulation: 1. The atmospheric circulation and the hydrology of the earth's surface. *Monthly Weather Review*, *97*(11), 739–774. https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2

Massey, F. J. Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78. https://doi.org/10.1080/01621459.1951.10500769

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, *51*, 716–728. https://doi.org/10.1002/2014WR015820

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of Hydrometeorology*, *17*(3), 745–759. https://doi.org/10.1175/JHM-D-15-0063.1

Niu, G.-Y., & Yang, Z.-L. (2004). Effects of vegetation canopy processes on snow surface energy and mass balances. *Journal of Geophysical Research*, *109*, D23111. https://doi.org/10.1029/2004JD004884

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. (2005). A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *Journal of Geophysical Research*, *110*, D21106. https://doi.org/10.1029/2005JD006111

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*, *112*, 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*, *116*, D12109. https://doi.org/10.1029/2010JD015139

Oleson, K. W., Dai, Y., Bonan, G., Bosilovich, M., Dickinson, R., Dirmeyer, P., et al. (2004). *Technical description of the community land model (CLM)*. Boulder, CO: National Center for Atmospheric Research.

Oleson, K. W., Niu, G.-Y., Yang, Z.-L., Lawrence, D. M., Thornton, P. E., Lawrence, P. J., et al. (2008). Improvements to the Community Land Model and their impact on the hydrological cycle. *Journal of Geophysical Research*, *113*, G01021. https://doi.org/10.1029/2007JG000563

Park, S., & Park, S. K. (2016). Parameterization of the snow-covered surface albedo in the Noah-MP Version 1.0 by implementing vegetation effects. *Geoscientific Model Development*, *9*(3), 1073–1085. https://doi.org/10.5194/gmd-9-1073-2016

Pilotto, I., Rodríguez, D., Tomasella, J., Sampaio, G., & Chou, S. (2015). Comparisons of the Noah-MP land surface model simulations with measurements of forest and crop sites in Amazonia. *Meteorology and Atmospheric Physics*, *127*(6), 711–723. https://doi.org/10.1007/s00703-015-0399-8

Pitman, A. J. (2003). The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology*, *23*(5), 479–510. https://doi.org/10.1002/joc.893

Sakumura, C., Bettadpur, S., & Bruinsma, S. (2014). Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. *Geophysical Research Letters*, *41*, 1389–1397. https://doi.org/10.1002/2013GL058632

Schaake, J. C., Koren, V. I., Duan, Q. Y., Mitchell, K., & Chen, F. (1996). Simple water balance model for estimating runoff at different spatial and temporal scales. *Journal of Geophysical Research*, *101*(D3), 7461–7475. https://doi.org/10.1029/95JD02892

Sellers, P. J., Dickinson, R. E., Randall, D. A., Betts, A. K., Hall, F. G., Berry, J. A., et al. (1997). Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. *Science*, *275*(5299), 502–509. https://doi.org/10.1126/science.275.5299.502

Seo, K. W., Wilson, C. R., Famiglietti, J. S., Chen, J. L., & Rodell, M. (2006). Terrestrial water mass load changes from Gravity Recovery and Climate Experiment (GRACE). *Water Resources Research*, *42*, W05417. https://doi.org/10.1029/2005WR004255

Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F., & Watkins, M. M. (2004). GRACE measurements of mass variability in the Earth system. *Science*, *305*(5683), 503–505. https://doi.org/10.1126/science.1099192

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

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99–114. https://doi.org/10.2307/3001913

Verseghy, D. L. (1991). CLASS—A Canadian land surface scheme for GCMs. I. Soil model. *International Journal of Climatology*, *11*(2), 111–133. https://doi.org/10.1002/joc.3370110202

Wang, P., Niu, G. Y., Fang, Y. H., Wu, R. J., Yu, J. J., Yuan, G. F., et al. (2018). Implementing dynamic root optimization in Noah-MP for simulating phreatophytic root water uptake. *Water Resources Research*, *54*, 1560–1575. https://doi.org/10.1002/2017WR021061

Wang, W., Bruyere, C., Duda, M., Dudhia, J., Gill, D., Kavulich, M., et al. (2017). *ARW version 3 modeling system user's guide*. Boulder, CO: National Center for Atmosphere Research.

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Brewer, M., et al. (2016b). 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*, 2750–2779. https://doi.org/10.1002/2015JD023733

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Kumar, S., et al. (2016a). 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*, 196–220. https://doi.org/10.1002/2015JD023889

Xia, Y., Mocko, D., Huang, M., Li, B., 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

Xia, Y., Sheffield, J., Ek, M. B., Dong, J., Chaney, N., Wei, H., et al. (2014). Evaluation of multi-model simulated soil moisture in NLDAS-2. *Journal of Hydrology*, *512*, 107–125. https://doi.org/10.1016/j.jhydrol.2014.02.027

Xia, Y., Yang, Z.-L., Stoffa, P. L., & Sen, M. K. (2005). Using different hydrological variables to assess the impacts of atmospheric forcing errors on optimization and uncertainty analysis of the CHASM surface model at a cold catchment. *Journal of Geophysical Research*, *110*, D01101. https://doi.org/10.1029/2004JD005130

Xue, B.-L., Wang, L., Yang, K., Tian, L., Qin, J., Chen, Y., et al. (2013). Modeling the land surface water and energy cycles of a mesoscale watershed in the central Tibetan Plateau during summer with a distributed hydrological model. *Journal of Geophysical Research: Atmospheres*, *118*, 8857–8868. https://doi.org/10.1002/jgrd.50696

Xue, Y., Sellers, P. J., Kinter, J. L., & Shukla, J. (1991). A simplified biosphere model for global climate studies. *Journal of Climate*, *4*(3), 345–364. https://doi.org/10.1175/1520-0442(1991)004<0345:ASBMFG>2.0.CO;2

Yang, K., He, J., Tang, W., Qin, J., & Cheng, C. C. (2010). On downward shortwave and longwave radiations over high altitude regions: Observation and modeling in the Tibetan Plateau. *Agricultural and Forest Meteorology*, *150*(1), 38–46. https://doi.org/10.1016/j.agrformet.2009.08.004

Yang, Z.-L., Cai, X., Zhang, G., Tavakoly, A. A., Jin, Q., Meyer, L. H., & Guan, X. (2011). *The community Noah land surface model with multi-parameterization options (Noah-MP)*. Austin, TX: The University of Texas at Austin.

Yang, Z.-L., & Dickinson, R. E. (1996). Description of the Biosphere-Atmosphere Transfer Scheme (BATS) for the soil moisture workshop and evaluation of its performance. *Global and Planetary Change*, *13*(1-4), 117–134. https://doi.org/10.1016/0921-8181(95)00041-0

Yang, Z.-L., & Niu, G.-Y. (2003). The versatile integrator of surface and atmosphere processes: Part 1. Model description. *Global and Planetary Change*, *38*(1–2), 175–189. https://doi.org/10.1016/S0921-8181(03)00028-6

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*, *116*, D12110. https://doi.org/10.1029/2010JD015140

Yao, Y., Liang, S., Cheng, J., Liu, S., Fisher, J. B., Zhang, X., et al. (2013). MODIS-driven estimation of terrestrial latent heat flux in China based on a modified Priestley–Taylor algorithm. *Agricultural and Forest Meteorology*, *171-172*, 187–202. https://doi.org/10.1016/j.agrformet.2012.11.016

Zhai, P., Zhang, X., Wan, H., & Pan, X. (2005). Trends in total precipitation and frequency of daily precipitation extremes over China. *Journal of Climate*, *18*(7), 1096–1108. https://doi.org/10.1175/JCLI-3318.1

Zhang, G., Chen, F., & Gan, Y. (2016). Assessing uncertainties in the Noah-MP ensemble simulations of a cropland site during the Tibet Joint International Cooperation program (JICA) field campaign. *Journal of Geophysical Research: Atmospheres*, *121*, 9576–9596. https://doi.org/10.1002/2016JD024928

Zheng, D., Van der Velde, R., Su, Z., Wen, J., Booij, M. J., Hoekstra, A. Y., & Wang, X. (2015). Under-canopy turbulence and root water uptake of a Tibetan meadow ecosystem modeled by Noah-MP. *Water Resources Research*, *51*, 5735–5755. https://doi.org/10.1002/2015WR017115

Zheng, D., Van Der Velde, R., Su, Z., Wen, J., & Wang, X. (2017). Assessment of Noah Land Surface Model with various runoff parameterizations over a Tibetan river. *Journal of Geophysical Research: Atmospheres*, *122*, 1488–1504. https://doi.org/10.1002/2016JD025572

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*, 14,387–14,402. https://doi.org/10.1002/2016JD025187

Zhou, L., & Huang, R. (2014). Regional differences in surface sensible and latent heat fluxes in China. *Theoretical and Applied Climatology*, *116*(3-4), 625–637. https://doi.org/10.1007/s00704-013-0975-0