

A performance-based multivariable weighting framework for WRF parameterization-scheme selection

Tianyu Gou^{1,2,3}, Yaoyang Deng^{1,2,3}, Jun Niu^{1,2,3}, and Shaozhong Kang^{1,2,3}

¹State Key Laboratory of Efficient Utilization of Agricultural Water Resources, Beijing 100083, China

²National Field Scientific Observation and Research Station on Efficient Water Use of Oasis Agriculture in Wuwei of Gansu Province, Wuwei 733009, China

³Center for Agricultural Water Research in China, China Agricultural University, Beijing 100083, China

Correspondence: Shaozhong Kang (kangsz@cau.edu.cn)

Abstract. The selection of physical parameterization schemes is a major source of uncertainty in WRF simulations. A comprehensive evaluation of model performance requires the simultaneous consideration of multiple output variables, but simple multivariable averaging can dilute the influence of variables that are difficult to simulate or that respond strongly to alternative physical-process options. To address this issue, this study proposes a performance-based multivariable weighting framework for WRF parameterization-scheme evaluation. The framework combines two components: mean relative error, which represents variable-level simulation difficulty quantified by relative simulation error, and the coefficient of variation of absolute errors, which represents the scheme-dependent response of each variable within the evaluated physical-process group. The approach was applied and evaluated in assessing WRF parameterization schemes across two climatically distinct environments: the arid Northwest and the humid coastal Southeast of China. The results show that the method assigns greater weights to variables that are difficult to simulate, such as precipitation and 10 m wind, thereby helping identify physically interpretable and well-performing parameterization-scheme combinations under the prescribed evaluation framework. Compared with equal weighting and established statistical weighting methods, the proposed $\text{MRE} \times \text{CV}$ weighting helped identify parameterization-scheme combinations that achieved the lowest integrated cMIEI in both representative screening years and in the independent 2010 extreme-year validation. These results indicate that jointly considering variable-level simulation difficulty and scheme-dependent response improves the diagnostic relevance of multivariable WRF parameterization-scheme evaluation.

1 Introduction

Earth system models (ESMs) support climate-change detection and attribution, future climate projection, and adaptation research (Reichler and Kim, 2008; Knutti et al., 2010). Their reliability depends on systematic evaluation against observations, especially as model representations of physical processes become broader and more interconnected (Reichstein et al., 2019; Eyering et al., 2019; Hausfather et al., 2020). Model evaluation commonly uses performance metrics to quantify how well simulations reproduce climate characteristics or processes (Baker and Taylor, 2016). For single variables, correlation, root mean square error, and variance ratio are widely used, and the Taylor diagram combines these statistics into a compact graphical summary (Lambert and Boer, 2001; Taylor, 2001). However, single-variable or scalar summaries are insufficient when model

performance must be assessed across multiple coupled variables (Pincus et al., 2008). Multivariable metrics such as the Model
25 Climate Performance Index (MCPI) and multivariate integrated evaluation (MVIE) therefore provide useful bases for integrated
model assessment (Gleckler et al., 2008; Xu et al., 2017).

The Weather Research and Forecasting (WRF) model is a regional climate model used within the Coordinated Regional
Downscaling Experiment (CORDEX) under the World Climate Research Programme (WCRP). It includes multiple parameter-
ization schemes for microphysics, land surface, planetary boundary layer, cumulus, and radiation processes (Skamarock et
30 al., 2019). Different schemes and combinations can produce large differences in simulated meteorological variables (Yu et al.,
2022b). Because physical processes interact and individual schemes are region- and variable-dependent, no universally optimal
WRF configuration can be assumed (Fernández et al., 2007; Borge et al., 2008). This range of selectable schemes is a major
source of WRF simulation uncertainty (Mooney et al., 2013), making region-specific parameterization-scheme evaluation an
important prerequisite for WRF applications (Chen et al., 2014; Ferreira et al., 2014). The size of the configuration space also
35 creates a practical constraint. In WRF version 4.3, the available schemes yield a theoretical upper bound of more than 14.5 mil-
lion possible combinations before physical compatibility constraints are considered (Wang et al., 2021). Exhaustive evaluation
is therefore impractical, and sensitivity analyses that vary schemes process by process are commonly used to identify suitable
configurations (Li et al., 2014; Yu et al., 2022a).

A remaining difficulty is how to evaluate WRF configurations when several meteorological variables are considered to-
40 gether. WRF produces multiple output variables, but evaluations often emphasize key or difficult-to-simulate variables such
as precipitation. Such univariate evaluation can bias model assessment by overlooking other variables that are important for
regional climate dynamics (Lv et al., 2020). Multivariable evaluation has therefore been increasingly used to assess simula-
tion performance more comprehensively and to represent parameterization-related uncertainty more transparently (Righi et al.,
2020; Dai et al., 2021; Lu et al., 2021). Yet many multivariable approaches still assign equal weights to all variables or rely
45 on subjective expert weighting. These choices can dilute variables with large simulation errors, such as precipitation and 10
m wind, within an averaged multivariable score. They can also ignore the fact that variables respond unevenly to alternative
physical-process schemes. For example, microphysics and cumulus schemes typically affect precipitation more strongly than
some other near-surface variables. Thus, WRF parameterization-scheme screening requires a weighting rule that can incorpo-
rate both variable-level simulation difficulty and scheme-dependent response.

50 To address this need, this study proposes a performance-based multivariable weighting framework for WRF
parameterization-scheme evaluation. The proposed $\text{MRE} \times \text{CV}$ weighting combines two components: mean relative error
(MRE), which represents variable-level simulation difficulty, and the coefficient of variation of absolute errors (CV), which
represents the scheme-dependent response of each variable within the evaluated physical-process group. By combining these
components, the framework gives greater influence to variables that are both difficult to simulate and informative for distin-
55 guishing candidate schemes. The method operates at the evaluation stage. It changes how variables contribute to the integrated
score and can therefore affect which parameterization-scheme combination is selected, but it does not directly modify the WRF
model physics or numerical simulations. We applied the method in two contrasting regions, the arid region of Northwest China

and the coastal region of Southeast China, and compared it with equal weighting, entropy weighting, coefficient-of-variation weighting, and CRITIC weighting. We further tested the selected configurations in an independent extreme climate year.

The remainder of this paper is organized as follows. Sect. 2 describes the study areas, model configuration, datasets, and implementation of the proposed weighting framework. Sect. 3 evaluates the parameterization combinations, identifies regionally optimal configurations, compares the proposed method with four alternative weighting strategies, and tests the selected schemes in an independent extreme climate year. Sect. 4 discusses the weight-allocation mechanism, applicability, and generalizability of the framework. Sect. 5 summarizes the main conclusions.

2 Method

2.1 Study area

This study selects the arid region of Northwest China and the coastal region of Southeast China as its research areas. This choice is primarily based on the significant differences between the two regions in terms of climate background, underlying surface characteristics, and dominant physical processes. The two regions exhibit distinct hydro-thermal patterns but share similarities and complementarities in the frequency of extreme climate events and the complexity of climate dynamics. The arid region of Northwest China is characterized by extreme temperature fluctuations and dust weather, whereas the coastal region of Southeast China is marked by intense precipitation events and frequent typhoons. These extreme climatic conditions present substantial challenges to the accuracy and reliability of regional climate models, making them ideal testbeds for evaluating model simulation capabilities under extreme climate scenarios. By independently applying the same evaluation methodology to these two regions, this study aims not only to verify the method's general capability to stably identify optimal parameterization schemes across different climate backgrounds but also to provide a critical scientific basis for future research on water resource management in arid regions and the development of rainstorm and flood early warning systems in coastal urban clusters.

2.1.1 Arid region of Northwest China

The arid region of Northwest China (73° – 106° E, 35° – 50° N) lies at the core of the Eurasian continent, encompassing the entire Xinjiang region, the Hexi Corridor in Gansu Province, and parts of Inner Mongolia, as illustrated in Fig. 1(b). The region features complex topography, including prominent mountain ranges such as the Kunlun Mountains, Qilian Mountains, and Tianshan Mountains, as well as extensive inland basins like the Tarim and Junggar Basins. Its deep inland location, remote from maritime moisture sources, coupled with its position within a subsiding air divergence zone north of the Tibetan Plateau, severely limits moisture influx from the ocean, resulting in an extremely arid climate. Annual precipitation typically falls below 200 mm, with hyper-arid zones such as the Tarim Basin and Turpan receiving less than 50 mm per year, making this one of the driest regions globally. While a few southeastern subregions exhibit characteristics of a temperate monsoon climate, the vast majority of the area experiences a temperate continental or alpine climate, marked by cold, dry winters, hot summers, low precipitation, high evaporation rates, and significant diurnal temperature variation. Owing to its geographical position and

distinctive climatic conditions, the region is highly sensitive to global climate change and regional anthropogenic activities—
 90 particularly regarding water cycle dynamics, temperature variability, and extreme climate phenomena such as droughts and
 heatwaves. This sensitivity underscores the region’s importance for understanding climate impacts in inland arid zones and
 across the Tibetan Plateau (Fu et al., 2022). Furthermore, due to its extreme aridity and climatic complexity, existing climate
 models often face challenges in accurately simulating regional climate change patterns. A comprehensive assessment and study
 of climate simulation performance in this region holds immense value for predicting climate change in arid areas globally.

95 **2.1.2 Coastal region of Southeast China**

In contrast to the arid region of Northwest, the coastal region of Southeast China is among the most precipitation-abundant areas
 in the country. Situated at the convergence of tropical and subtropical climatic zones, the region features a diverse landscape
 of coastal plains, mountains, and hills. It is also one of China’s most densely populated and highly urbanized regions, as
 illustrated in Fig. 1(c), and serves as a major hub of economic activity. Heavily influenced by the southeast monsoon and
 100 frequent typhoons, the region benefits from a substantial influx of warm, moist air, resulting in annual precipitation exceeding
 1,500 mm across most areas. The coastal region of Southeast is characterized by relatively high temperatures, minimal diurnal
 temperature variation, and a general equilibrium between precipitation and evaporation. This stands in sharp contrast to the
 arid, rainfall-deficient Northwest Arid Region, where large temperature fluctuations dominate. However, under the influence of
 global climate change and accelerated urbanization, the frequency and intensity of extreme climate events, such as typhoons,
 105 heavy rainfall, and droughts, have increased markedly in this region (Yang et al., 2015; Li et al., 2021), posing serious risks to
 both human populations and ecological systems. Accordingly, a comprehensive investigation into the regional climate system,
 particularly the coupled interactions among oceanic, atmospheric, and terrestrial components, is essential for enhancing climate
 prediction accuracy, mitigating the impacts of extreme events, and fostering sustainable regional development.

2.2 Model configuration and datasets

110 In this study, the Advanced Research WRF model version 4.3.3 was used for numerical simulations. Different model config-
 urations were set for the two study regions (Table 1). For the arid region of Northwest China, the simulation domain covers
 30°N to 52°N and 66°E to 114°E, with a horizontal grid resolution of 9 km and a total of 380 × 240 grid points. The model had
 38 vertical levels, with the top of the model at 50 hPa. For the coastal region of Southeast China, the simulation domain covers
 20°N to 29°N and 108°E to 120°E, with the same horizontal grid resolution of 9 km and a total of 120 × 105 grid points. The
 115 vertical configuration was identical to that of the arid region of Northwest.

Table 1: Overview of WRF model configurations.

	Northwest China	Southeast China
Model version	WRF/ARW (4.3.3)	WRF/ARW (4.3.3)

Horizontal resolution	9km	9km
Horizontal grid	380 (E–W) × 240 (S–N)	120 (E–W) × 105 (S–N)
Initial and boundary conditions	ERA5	ERA5
Evaluation period	June 1st–Aug 31st, 2007	June 1st–Aug 31st, 2009
Spin-up period	May 1st–May 31st	May 1st–May 31st
Land-use data	GLC_FCS30D	GLC_FCS30D
Default parameterization schemes	Microphysics–Thompson	Microphysics–Thompson
	PBL/surface layer scheme–MYJ	PBL/surface layer scheme–MYJ
	Cumulus–Tiedtke	Cumulus–Tiedtke
	Radiation–RRTMG	Radiation–RRTMG
	LSM–Noah LSM	LSM–Noah LSM

For this study, the period from May 1 to August 31 was selected for 2007 in the arid region of Northwest China and 2009 in the coastal region of Southeast China to screen for optimal parameterization scheme combinations using both the proposed MRE × CV weighting and equal weighting methods. The first month, from May 1 to May 31, was used as the model spin-up period, and only the June–August simulations were used to evaluate the performance of the physical parameterization scheme combinations. The selection of simulation years for the two study regions was based on long-term meteorological station observations from 1991 to 2020. For each candidate year, the observed multivariable meteorological conditions were compared with the corresponding 30-year climatological reference using centered multivariable integrated evaluation (MVIE) statistics, including the centered root-mean-square vector difference, vector similarity coefficient, standard deviation ratio, and cMIEI. The cMIEI provides an integrated measure of the overall distance between each candidate year and the climatological reference, while the variable-level metrics were used to check whether the selected years were dominated by anomalous deviations in individual variables. The analysis showed that 2007 and 2009 had the lowest cMIEI values in their respective regions and did not show pronounced outlying deviations among the seven variables evaluated in the WRF assessment. Overall, these two years ranked among the candidate years closest to the long-term climatological state across the multivariable and variable-level metrics (see Tables S2–S3 and Figs. S1–S2 in Supporting Information).

As a single independent extreme-year validation, the year 2010, which was characterized by extreme climate events common to both study regions, was selected to examine whether the schemes obtained from the screening period retained their evaluated performance under a different climate background. The optimal scheme obtained from the screening period was applied to the validation period for simulation, and its performance was subsequently evaluated. The year 2010 is a representative year of extreme climate in China, during which the arid region of Northwest experienced drought and the coastal region of Southeast

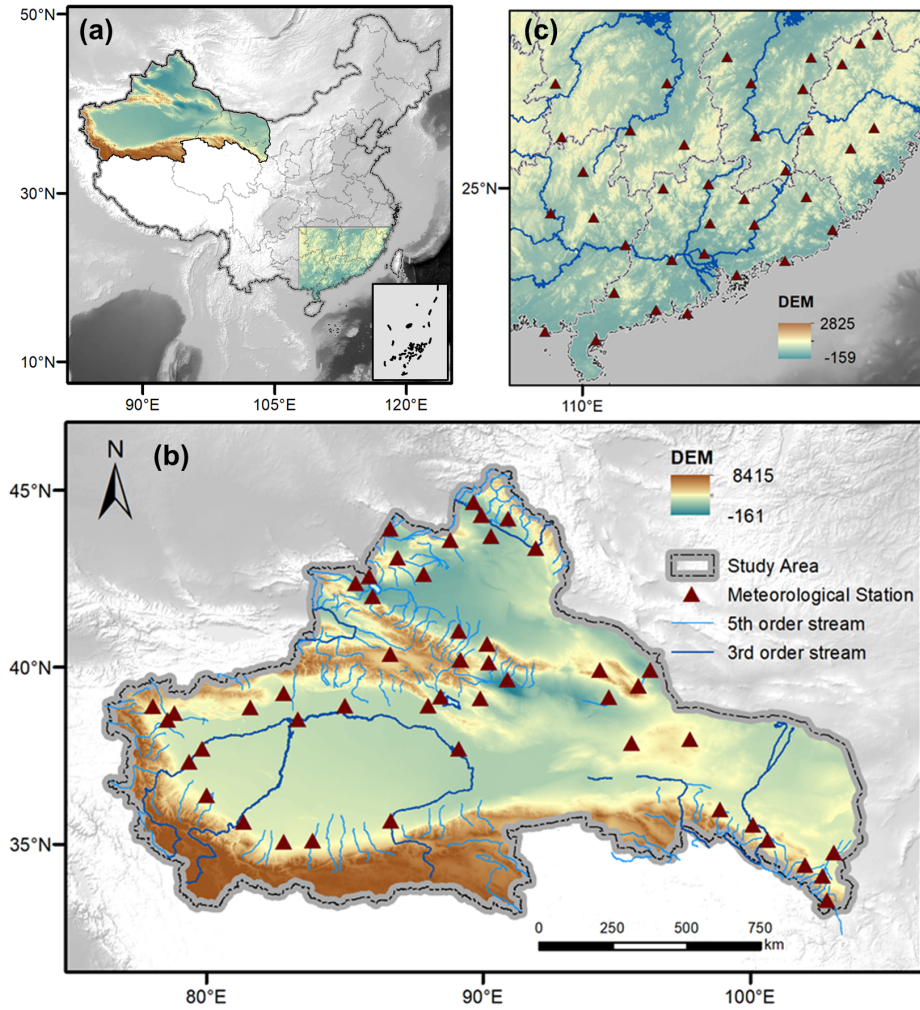


Figure 1. (a) Geographical location of the study area. (b) Overview of the arid region of Northwest China. (c) Overview of the coastal region of Southeast China.

135 was affected by severe flooding. From June to July in the Northwest, persistent high temperatures and markedly reduced rainfall, 30% to 80% below average, were observed. Concurrently, southern China experienced heavy rainfall in June and August, leading to severe flooding in the coastal region of Southeast, with substantial loss of life and property. Therefore, the 2010 experiment should be interpreted as an independent test under one extreme climate year, rather than as a temporal cross-validation or a proof of multi-year robustness.

140 To reduce the computational burden associated with the high-dimensional WRF physics-configuration space, this study adopted a pre-specified conditional screening strategy. The physical processes were evaluated sequentially in the order of mi-

145 crophysics, planetary boundary layer, cumulus, radiation, and land surface processes, and the best-performing parameterization
 scheme selected at each step was carried forward as the background configuration for the subsequent step. This order follows
 procedural precedents in previous WRF parameterization sensitivity and physics-ensemble evaluation studies, including the
 stepwise decision approach of Stergiou et al. (2017), which begins with microphysics and then evaluates PBL, cumulus, ra-
 150 diation, and land surface scheme groups, and the WRF physics-ensemble evaluation of Dai et al. (2021). It is also consistent
 with empirical evidence from WRF ensemble evaluations and the present results that microphysics, planetary boundary layer,
 and cumulus schemes often exert strong influences on summer near-surface meteorological performance, particularly for pre-
 cipitation, 10 m wind, temperature, and humidity, which are key variables in the present multivariable evaluation. Therefore,
 155 the prescribed sequence is used here as a computationally feasible stepwise screening heuristic, rather than as a theoretically
 unique physical hierarchy or a guarantee of global optimality. For this study, a near-complete ensemble of the parameterization
 schemes available within the WRF model was utilized, as presented in Table 2. The final number of scheme combinations was
 constrained by incompatibilities identified between certain physical process parameterizations. This resulted in 42 (proposed
 MRE $\times$ CV weighting method) /43 (equal weighting method) combinations for the arid domain in Northwest China and 54
 (proposed MRE $\times$ CV weighting method) /53 (equal weighting method) combinations for the coastal domain in Southeast
 China (see Tables S4-S5 and S6-S7 in Supporting Information).

Table 2: Physics parameterization scheme assessed in this study.

	Microphysics	Planetary Boundary Layer/Surface Layer	Cumulus	Radiation	Land surface
1	8 Thompson	2 MYJ+2 MYJ	6 Tiedtke	1 RRTM+1 Dudhia	2 Noah LSM
2	1 Kessler	1 YSU+1 MM5	1 KF	3 CAM+3 CAM	1 SLAB
3	2 Lin	3 GFS+3 GFS	2 BMJ	5 New Goddard	4 NoahMP
4	3 WSM3	4 QNSE+4 QNSE	3 GF	99 GFDL+99 GFDL	5 CLM4
5	4 WSM5	5 MYNN2.5+1 MM5	5 Grell3		
6	5 Eta/Ferrier	5 MYNN2.5+2 MYJ	7 ZM		
7	6 WSM6	5 MYNN2.5+5 MYNN	11 MsKF		

8	9 Milbrandt	6 MYNN3+ 1 MM5	14 NSAS
9	10 Morrison	6 MYNN3 + 2 MYJ	16 New Tiedtke
10	11 CAM	6 MYNN3 + 5 MYNN	93 GD
11	13 Lin-SBU	7 ACM2+1 MM5	99 Old KF
12	14 WDM5	7 ACM2+7 Pleim-Xiu	
13	16 WDM6	8 BouLac + 1 MM5	
14	17 NSSL 2-17	8 BouLac + 2 MYJ	
15	18 NSSL 2-18	9 UW + 1 MM5	
16	19 NSSL 1-19	9 UW + 2 MYJ	
17	21 NSSL 1-21	11 Shin-Hong+1 MM5	
18	24 WSM7	12 GBM + 1 MM5	
19	26 WDM7	99 MRF+1 MM5	
20	28 Thompson aerosol-aware		
21	50 P3		

The model's initial and boundary conditions were provided by the ECMWF's fifth-generation atmospheric reanalysis dataset (ERA5), with a spatial resolution of $0.25^{\circ} \times 0.25^{\circ}$ and a temporal resolution of 6 h. The reanalysis dataset is available on the European Center for Medium-Range Weather Forecasting (<https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-l>
160 evels ; <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels>).

Owing to the lack of ground validation and limitations in spatial and temporal resolution, the default MODIS land use dataset in the WRF model exhibits discrepancies with actual land use results in the study areas. This study uses the 1985–2022 global 30-m land use product (GLC_FCS30D) developed by the Aerospace Information Research Institute, Chinese Academy of Sciences, as the surface input data (Zhang et al., 2024), available for download from <https://zenodo.org/records/8239305>.

165 This dataset meets the requirements for continuous and consistent land use data in climate modeling, with a spatial resolution

of 30 m and a time span from 1985–2022, making it suitable for more detailed regional simulation studies. The land use types in the dataset were reclassified according to the IGBP20 classification standard.

To comprehensively assess the performance of the WRF model, observational data from meteorological stations were used to validate the simulated variables. Daily meteorological observation data were obtained from the China Meteorological Data Service Centre (<https://data.cma.cn/en>). A total of 49 meteorological stations from the arid region of Northwest China and 38 from the coastal region of Southeast China, which provided long-term daily meteorological data from 1991–2020, were collected. The spatial distribution of the meteorological stations in the study area is shown in Fig. 1. These station-based observations underwent rigorous calibration and quality control, and thus provide a more accurate depiction of surface meteorological conditions than gridded datasets. The seven evaluated variables were 2 m air temperature (TEM), ground surface temperature (GST), air pressure (PRS), precipitation (PRE), relative humidity (RHU), and the zonal (U10) and meridional (V10) components of 10 m wind.

Variable weighting uses six evaluation groups: GST, PRE, PRS, RHU, TEM, and UV10. Following the vector-field evaluation framework (Xu et al., 2016; Zhang et al., 2021), UV10 denotes the two-dimensional 10 m wind vector (U10, V10), evaluated jointly under one group weight rather than converted to wind-speed magnitude. U10 and V10 remain separate components in the variable-level scatter plots and tables.

2.3 Model evaluation and variable weighting method

2.3.1 Taylor diagram and related statistical metrics

This study utilizes several multivariate evaluation methods to describe the performance of the WRF model. The normalized Taylor diagram (Taylor, 2001) is used to assess the performance of different physical schemes in the sensitivity experiments to find the optimal combination of WRF schemes. The Taylor diagram simultaneously displays three complementary statistical metrics related to model performance: the center root mean square error (cRMSE), standard deviation (σ), and correlation coefficient (R). In addition, the vector statistical method (Xu et al., 2016) is employed, allowing the statistical metrics to reflect the combined results of multiple variables.

$$\sigma_{fn} = \sqrt{\sum_{m=1}^M \omega_m \left(\frac{1}{T} \sum_{t=1}^T (f_{mnt} - \overline{f_{mn}})^2 \right)} \quad (1)$$

$$\sigma_r = \sqrt{\sum_{m=1}^M \omega_m \left(\frac{1}{T} \sum_{t=1}^T (r_{mt} - \overline{r_m})^2 \right)} \quad (2)$$

$$R_n = \frac{\sum_{m=1}^M \omega_m \left(\frac{1}{T} \sum_{t=1}^T (f_{mnt} - \overline{f_{mn}}) (r_{mt} - \overline{r_m}) \right)}{\sigma_{fn} \sigma_r} \quad (3)$$

$$cRMSE_n = \sqrt{\sum_{m=1}^M \omega_m \left(\frac{1}{T} \sum_{t=1}^T ((f_{mnt} - \overline{f_{mn}}) - (r_{mt} - \overline{r_m}))^2 \right)} \quad (4)$$

In the equation, m indexes the variable group, and n represents the model's parameterization scheme. In the WRF parameterization-scheme evaluation, t denotes the valid daily station-level sample after pooling the station and day dimensions. The screening-year evaluations used 92 June-August daily values at 49 stations in Northwest China and 38 stations in Southeast China, after applying the common missing-value mask. The same daily station-level aggregation was used for R, cRMSE, cMIEI, and the Taylor diagrams. These metrics were not calculated from monthly means, seasonal means, or climatological averages. f_{mnt} represents the simulated value of the m variable at the t -th valid station-day sample for the model's n parameterization scheme, whereas r_{mt} represents the corresponding observed value. ω_m represents the weight of the m variable during the statistical calculation. For UV10, the products and squares in Eqs. (1)–(4) are interpreted as vector inner products and squared Euclidean norms, respectively, with each wind component centered separately.

The five methods' final configurations were compared using unweighted RMSE, MAE, mean bias, and Pearson correlation for all seven variables, with common valid daily station samples across methods and variables within each region-year (Table S15).

2.3.2 Model performance indicators

Although Taylor diagrams provide a multidimensional perspective on model performance through multiple statistical indicators, they are limited in their ability to fully quantify model superiority in complex evaluation scenarios. While it is widely recognized that a single performance metric cannot capture all aspects of model evaluation across multiple dimensions, the establishment of a unified evaluation standard is essential for this study to more intuitively illustrate the process of parameterization scheme selection. To this end, the multivariable integrated evaluation index (MIEI) (Zhang et al., 2021) is adopted as the primary metric for assessing model simulation performance.

MIEI denotes the multivariable integrated evaluation index. This study uses its centered form, denoted cMIEI, in which the simulated and observed series are each centered by subtracting their own mean over the evaluation samples. The index combines agreement in the amplitude and pattern of the centered fields. For a given evaluation set and variable weights, lower cMIEI indicates better agreement. The multivariable evaluation framework and its applications are described by Xu et al. (2016, 2017), Lv et al. (2020), and Zhang et al. (2021). In the WRF evaluation, cMIEI and the Taylor-diagram statistics use the same daily station-level June-August samples; the 30-year climatological reference is used only for representative-year selection. The centered index used here is

$$cMIEI_n^2 = \sum_{m=1}^M \omega_m (L_{mn}^* - 1)^2 + 2(1 - R_n) \quad (5)$$

$$220 \quad L_{mn}^* = \frac{\sigma_{fmn}}{\sigma_{rm}} \quad (6)$$

Here, m indexes the six variable groups, n indexes a candidate configuration, and $\sum_{m=1}^M \omega_m = 1$. For a scalar variable, σ_{fmn} and σ_{rm} are the simulated and observed standard deviations. For UV10, they are the centered root mean square lengths of the two-component wind vectors. L_{mn}^* is their direct ratio. R_n is the centered vector similarity coefficient (cVSC) of the integrated fields, calculated after scaling each variable group by its observed root mean square length and applying the variable weights. The single-group form gives the variable-level cMIEI values. The prefix c denotes mean removal.

2.3.3 Performance-based multivariable weighting framework

The performance-based multivariable weighting framework proposed in this study is used to calculate the variable weight ω_m in the evaluation metrics described above. The term performance-based is used because both weighting components are derived from the simulated performance of candidate WRF parameterization schemes against observations. This framework does not update weights online during WRF integration. Instead, the weights are recalculated after the simulations for each candidate physical-process group are completed and before the candidate schemes are ranked. The framework follows a simple evaluation principle: if a variable is difficult to simulate and its error responds strongly to the candidate schemes in the evaluated physical-process group, that variable should receive greater influence in the integrated evaluation. The two components are the mean relative error ($\overline{MRE}_m$) and the coefficient of variation of absolute errors ($CV_{|ME|_m}$). The MRE and absolute-error CV components are therefore calculated on the same daily station-level model-observation pairs before being summarized across candidate schemes within each physical-process group.

The mean relative error ($\overline{MRE}_m$) represents variable-level simulation difficulty, quantified by the relative magnitude of simulation error across the candidate schemes in the evaluated physical-process group. Larger MRE values indicate variables for which the candidate schemes show weaker simulation ability and larger deviations from observations.

$$240 \quad \overline{MRE}_m = \frac{1}{N} \sum_{n=1}^N MRE_{mn} \quad (7)$$

$$= \frac{1}{N} \sum_{n=1}^N \left(\frac{|ME|_{mn}}{\frac{1}{T} \sum_{t=1}^T |r_{mt}|} \right) \quad (8)$$

$$= \frac{1}{N} \sum_{n=1}^N \left(\frac{\frac{1}{T} \sum_{t=1}^T |f_{mnt} - r_{mt}|}{\frac{1}{T} \sum_{t=1}^T |r_{mt}|} \right) \quad (9)$$

In the equation, f_{mnt} represents the simulated values of the m variable at the t discrete points for the model's n parameterization scheme, while r_{mt} represents the observed values of the m variable at the t discrete points. $|ME|_{mn}$ represents the absolute error of the m variable in the n parameterization scheme of the model.

The coefficient of variation of the absolute error ($CV_{|ME|m}$) represents the scheme-dependent response of each variable within the evaluated physical-process group, quantified by the variability of absolute errors among candidate schemes. A larger value indicates that the variable is more useful for distinguishing alternative parameterization options in that evaluation stage.

$$CV_{|ME|m} = \frac{\sqrt{\frac{1}{N} \sum_{n=1}^N \left(|ME|_{mn} - \overline{|ME|_m} \right)^2}}{\overline{|ME|_m}} \quad (10)$$

$$= \frac{\sqrt{\frac{1}{N} \sum_{n=1}^N \left(\frac{1}{T} \sum_{t=1}^T |f_{mnt} - r_{mt}| - \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{T} \sum_{t=1}^T |f_{mnt} - r_{mt}| \right) \right)^2}}{\frac{1}{N} \sum_{n=1}^N \left(\frac{1}{T} \sum_{t=1}^T |f_{mnt} - r_{mt}| \right)} \quad (11)$$

In the equation, f_{mnt} represents the simulated values of the m variable at the t discrete points for the model's n parameterization scheme, whereas r_{mt} represents the observed values of the m variable at the t discrete points. $|ME|_{mn}$ represents the absolute error of the m variable in the n parameterization scheme of the model. $\overline{|ME|_m}$ indicates the absolute error of the m variable for all parameterization schemes of the model.

The base of the method is to assign higher weights to variables with larger relative errors (poor simulation performance) and higher coefficients of variation (stronger scheme-dependent response within the evaluated physical-process group). We used the Softmax transformation to convert each statistical vector across variables into a set of positive weights that sum to one. This choice is consistent with the purpose of the proposed evaluation framework: variables that are difficult to simulate or whose errors respond strongly to alternative parameterization schemes should not be diluted in a multivariable average. At the same time, because Softmax assigns a positive value to every variable, no variable is removed from the integrated evaluation. The use of Softmax is therefore a methodological choice for balancing two requirements: preserving the multivariable nature of the evaluation while giving greater influence to limiting variables.

To make the normalization explicit, the two component weights are calculated over the M output variables as follows:

$$\omega_{m1} = \frac{e^{\overline{MRE}_m}}{\sum_{j=1}^M e^{\overline{MRE}_j}} \quad (12)$$

$$\omega_{m2} = \frac{e^{CV_{|ME|m}}}{\sum_{j=1}^M e^{CV_{|ME|j}}} \quad (13)$$

$$\omega_m = \frac{\omega_{m1}\omega_{m2}}{\sum_{j=1}^M \omega_{j1}\omega_{j2}} \quad (14)$$

In the equation, ω_{m1} and ω_{m2} represent the component weights of variable m derived from the mean relative error and the coefficient of variation of the absolute error, respectively. The final variable weight ω_m is obtained by multiplying the two

component weights and then normalizing the product across all variables. This second normalization ensures that $\sum_{m=1}^M \omega_m = 1$, as required by the multivariable evaluation metrics.

To isolate the contribution of the absolute-error CV component, we additionally evaluated an MRE-only baseline, $\omega_m^{\text{MRE}} = \exp(\overline{MRE}_m) / \sum_{j=1}^M \exp(\overline{MRE}_j)$. At each evaluation stage, this baseline and the proposed rule used the same candidate simulations, observations, variable grouping, sample handling, and cMIEI formulation; only the variable weights were changed. We compared both the variable weights and the candidate-scheme rankings and checked whether the selected scheme changed at any stage of the sequential pathway.

To examine whether the weight distribution depends on the specific normalization function, we also conducted a sensitivity analysis using linear normalization and temperature-scaled Softmax transformations. The temperature-scaled Softmax is defined as $w_m(\tau) = \exp(s_m/\tau) / \sum_{j=1}^M \exp(s_j/\tau)$ for a statistical score s_m . The original formulation corresponds to $\tau = 1$; smaller values of τ produce more concentrated weights, whereas larger values produce smoother weight distributions. As a separate sensitivity test of the component-combination rule, we also calculated an equally weighted additive alternative:

$$\omega_m^{\text{add}} = \frac{\alpha \omega_{m1} + (1 - \alpha) \omega_{m2}}{\sum_{j=1}^M [\alpha \omega_{j1} + (1 - \alpha) \omega_{j2}]}, \quad \alpha = 0.5. \quad (15)$$

Both combination rules allow compensation between MRE and CV. Substituting Eqs. (12) and (13) into Eq. (14) shows that the product rule is equivalent to applying Softmax to $\overline{MRE}_m + CV_{|ME|m}$ across variables, whereas Eq. (15) averages the two separately normalized component weights. The selected-scheme consequences of this alternative rule are reported in Table S10.

Together, the selection of the MRE and CV components, the Softmax transformation, and the multiplicative combination define a reproducible evaluation rule under explicit methodological choices. The framework is therefore intended to make variable weighting transparent and testable, rather than to claim a unique objective set of variable weights.

We further evaluated whether the derived weights were sensitive to the station sample, the summer time period, or the evaluated variable set. For each region and physical-process evaluation stage, stations were resampled with replacement 500 times while retaining all summer days. Temporal stability was evaluated using 500 moving-block bootstrap samples composed of contiguous 14 d blocks, together with separate June, July, and August subsets. The MRE component, absolute-error CV component, and final normalized weights were recalculated for every resample. Each resulting weight vector was compared with the full-sample vector using the Spearman rank correlation, maximum absolute weight difference, and retention rates of the highest-weight variable and the two highest-weight variables. Variable-set stability was evaluated analytically and numerically by deleting each variable group in turn and comparing the recomputed weights of the retained variables with their theoretical renormalization. Full definitions and results are provided in Fig. S6 and Tables S8-S9.

We also examined the ranking stability of the selected stage-level parameterization schemes. For each region-stage comparison, 500 station-bootstrap samples and 500 14 d moving-block-bootstrap samples were used to recalculate the proposed weights, cMIEI values, and candidate-scheme rankings. The selected scheme was compared with nearby competitors using 95% bootstrap intervals of ΔcMIEI , and its top-1, top-2, and top-3 rank retention rates were summarized in the Supporting

Information (Fig. S8 and Table S14). Because physically related meteorological variables can carry overlapping information, we also diagnosed inter-variable dependence in two data views. Pearson correlation matrices were calculated from the observed daily station series and from the scheme-level standardized-error matrices used in the proposed screening workflow. We then
305 performed a targeted no-GST sensitivity test while retaining TEM as the standard near-surface air-temperature variable, to examine whether the strongest temperature-related dependence dominated the selected stage-level schemes. These diagnostics are reported in Fig. S7 and Table S12.

The specific application process of this method in WRF model evaluation includes the following steps:

Step 1: Run all the parameterization schemes in the first physical parameterization process (e.g., microphysics schemes) and
310 calculate the absolute and relative errors of the simulation results for different variables compared with the observed values.

Step 2: Calculate the mean relative error and the coefficient of variation of the absolute error for each variable to quantify variable-level simulation difficulty and the scheme-dependent response of variable errors within the evaluated physical-process group.

Step 3: Use the Softmax function to convert the two statistical indicators (mean relative error and coefficient of variation) into
315 two component weights, multiply the two components, and normalize the product across variables to obtain the final variable weights used in the parameterization-scheme evaluation.

Step 4: Based on the obtained weights, further calculate the statistical indicators for each scheme, such as the RMSE, R, σ , and cMIEI, to determine the best scheme for the first physical parameterization process.

Step 5: Use the best scheme identified in Step 4, continue simulating the next physical parameterization process and repeat
320 the above steps until the best scheme for all physical parameterization processes is determined.

2.3.4 Benchmark weighting methods

To examine how the proposed weighting scheme differs from established statistical weighting approaches, we included entropy weighting, coefficient-of-variation weighting, and CRITIC weighting as benchmark methods. Equal weighting was retained as a simple reference baseline. All five methods used the same candidate WRF parameterization outputs, the same station
325 observations, the same evaluation variables, and the same MVIE/cMIEI evaluation framework. All evaluation metrics in this comparison were calculated from daily station-level June-August model-observation pairs. Therefore, differences among the methods mainly reflect differences in how the variable weights were calculated, rather than differences in model input data or evaluation metrics.

Entropy weighting assigns larger weights to variables that contain more discriminating information across the candidate
330 schemes. For each variable, the scheme-level error values were first converted into normalized proportions, and the entropy value was then used to measure the degree of dispersion in the error distribution. Variables with lower entropy provide stronger differentiation among candidate schemes and therefore receive larger weights after normalization. Coefficient-of-variation weighting uses the ratio between the standard deviation and the mean error across candidate schemes. This method also emphasizes variables with larger relative dispersion, but its weighting response is directly controlled by the ratio of the spread to
335 the mean error. CRITIC weighting combines two components: the contrast intensity of a variable, represented by the standard

deviation of the normalized error values, and the conflict among variables, represented by the correlation structure among the variables. A variable receives a higher CRITIC weight when it varies strongly among candidate schemes and contributes information that is less redundant with other variables.

The proposed $MRE \times CV$ weighting differs from these general statistical weighting methods in its evaluation target. It is designed for WRF parameterization-scheme screening, where a variable should receive greater influence when it is both difficult to simulate and responsive to alternative parameterization choices within the evaluated physical-process group. The MRE component therefore represents variable-level simulation difficulty, whereas the absolute-error CV component represents the scheme-dependent response of that variable within a given physical-process group. The final weight combines these two components and can therefore be interpreted as a performance-based multivariable evaluation weight rather than only a measure of statistical dispersion. A conceptual comparison of the five weighting strategies is provided in Table S1.

3 Result

3.1 Optimal parameterization scheme combination for the arid region of Northwest

Figure 2 shows the weights assigned to variables in the WRF simulations of different physical processes for the arid region of Northwest using the proposed $MRE \times CV$ weighting method. Figure 2(a) shows the weights ω_{m1} assigned based on the mean relative error ($\overline{MRE_m}$), where 10 m wind and precipitation received average weights of 0.45 and 0.24, respectively, while all other variables had weights below 0.10. The higher weights for 10 m wind and precipitation indicate that all physical processes in the WRF model have large errors in simulating these variables, reflecting the model's weaker simulation capability for these variables. Therefore, more attention needs to be given to these variables in the model.

Figure 2(b) shows the weights ω_{m2} assigned based on the coefficient of variation ($CV_{|ME|m}$) of the absolute error, where the weights for different variables showed considerable variation during the simulation process, reflecting differences in the scheme-dependent response of variable errors within each physical-process evaluation stage. Precipitation and temperature were the variables exhibiting the most pronounced variability. Precipitation carried weight values of 0.17 and 0.18 in the micro-physics and cumulus processes, respectively, indicating a strong response to alternative schemes in these two physical-process groups. In contrast, temperature showed a stronger response in the radiation and land surface processes, with corresponding weight values of 0.184 and 0.201. Relative humidity showed a more pronounced response during the planetary boundary layer process, with a weight value of 0.176.

Figure 3(a) displays the median-relative cMIEI results for evaluating WRF parameterization schemes in the arid region of Northwest, providing an assessment of simulation performance across various output variables and in the overall integrated evaluation. Analysis of the simulation results for individual variables revealed substantial variability in precipitation simulation performance among different parameterization schemes. Notably, the P3 and MYJ schemes exhibited superior performance in simulating precipitation, as indicated by their lower cMIEI values, which reflected greater accuracy. In contrast, the Kessler and CLM schemes yielded higher cMIEI values, suggesting larger simulation errors. Although CLM, as a land surface model, does not directly govern precipitation-related processes, its influence on the hydrological cycle and land-atmosphere interactions

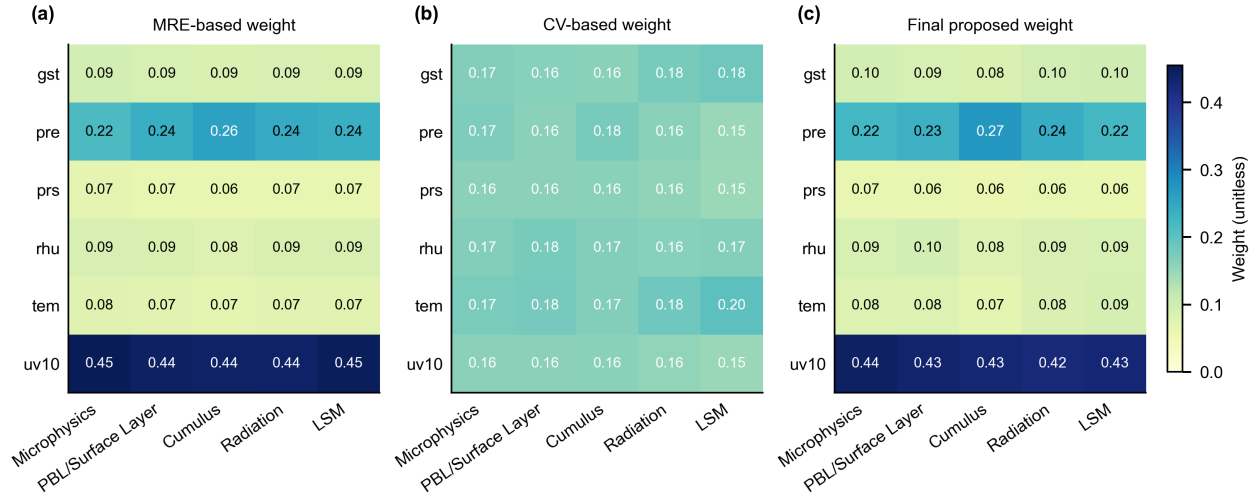


Figure 2. Weight distributions of the variables associated with the different physical processes of the WRF simulation in the arid region of Northwest. (a) MRE-based weight, (b) CV-based weight, and (c) final proposed weight. Weights are unitless.

affects precipitation outcomes. In the simulations of surface and air temperature, the impact of specific physical processes on these variables is more direct. For example, the Noah-MP and CLM4 schemes achieved the lowest and highest cMIEI values, respectively, for land surface processes, underscoring the critical role of land surface parameterizations in air temperature simulation. Additionally, the YSU scheme demonstrated strong performance in simulating 10 m wind within the planetary boundary layer process and was subsequently selected as the optimal scheme for further assessment. Consequently, the subsequent parameterization schemes for cumulus, radiation, and land surface processes also demonstrated good performance for 10 m wind.

In the multivariate integrated evaluation (rightmost column), the cMIEI value serves as the primary criterion for selecting parameterization schemes in this study. As the evaluation of physical processes advances, a progressive decrease in cMIEI values was observed, indicating that the sequentially selected configurations achieved lower integrated evaluation scores. The integrated evaluation identified the P3 (microphysics), YSU (planetary boundary layer), Tiedtke (cumulus), GFDL (radiation), and Noah LSM (land surface) schemes as the optimal choices indicating that their combination yielded the most suitable configuration for the arid region of Northwest. However, it is important to note that the optimal parameterization scheme does not necessarily deliver the best performance across all individual variables. For instance, within the radiation process, the GFDL scheme exhibited slightly inferior performance compared to the RRTMG scheme in simulating relative humidity and precipitation. Nonetheless, its stronger evaluated performance for 10 m wind and temperature contributed to a lower overall cMIEI value, establishing GFDL as the relatively best-performing scheme in the integrated evaluation.

Compared to cMIEI, the Taylor diagram offers a more comprehensive statistical evaluation of model performance, highlighting the discrepancies between different parameterization schemes and observational data. Figure 3(b-f) present the centered

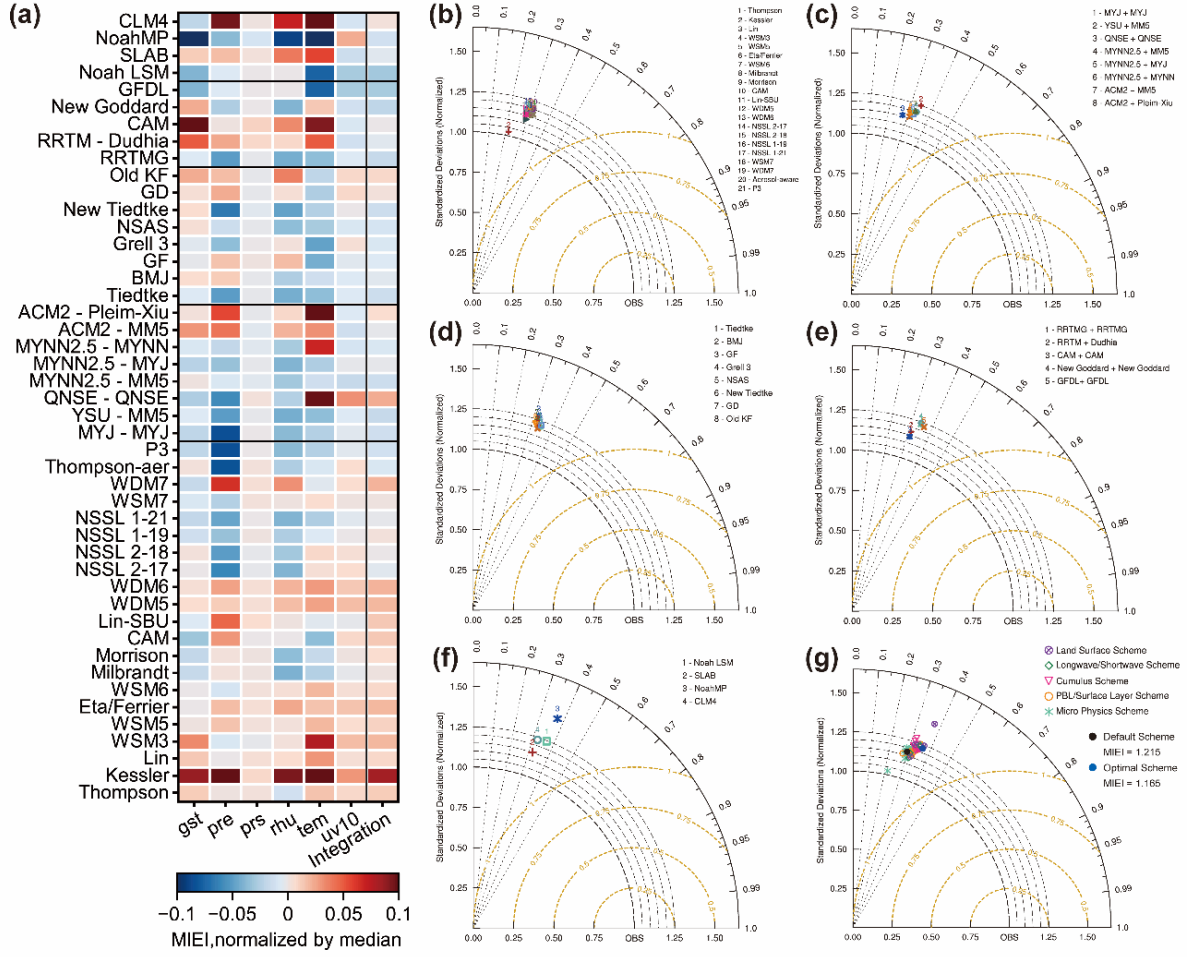


Figure 3. Evaluation results of parameterization schemes in the arid region of Northwest. (a) Median-relative cMIEI, $(cMIEI - C_{med})/C_{med}$, where C_{med} is the regional reference median calculated separately for each variable column and the integrated column. This transformation is for visualization only; negative values indicate lower cMIEI. Colors saturate at ± 0.10 . (b-f) Taylor diagrams of different physical processes. (b) Microphysics, (c) Planetary boundary layer, (d) Cumulus, (e) Radiation, (f) Land surface, and (g) Taylor diagrams of the candidate configurations pooled across all five physical-process stages. All cMIEI and Taylor-diagram statistics were calculated from daily station-level June-August values.

root mean square error (cRMSE), standard deviation (σ), and correlation coefficient (R) for each scheme across the various physical processes. In the microphysics process, the optimal parameterization scheme, P3, exhibited a relatively low RMSE of 1.232 and the highest correlation coefficient of 0.321. Most microphysics schemes demonstrated elevated standard deviation ratios (ranging from 1.156 to 1.205), indicating a tendency to overestimate the amplitude of anomalies in the simulated fields. For the planetary boundary layer process, the YSU scheme emerged as optimal, with the highest R value (0.346). Although

YSU performed well in terms of correlation, it still overestimated the amplitude of anomalies, as reflected in the standard deviation. In the cumulus process, the Tiedtke scheme demonstrated superior performance, with the lowest RMSE (1.277). For the radiation process, the GFDL scheme yielded the highest correlation coefficient (0.363). Although the CAM scheme showed better performance in terms of RMSE and standard deviation, its lower correlation coefficient (0.312) led to a lower integrated evaluation score compared to GFDL. In the land surface process, although the Noah LSM scheme did not achieve the best results in RMSE (1.282), correlation coefficient (0.365), its overall balanced performance across all indicators rendered it the most suitable parameterization scheme in the integrated evaluation.

The bootstrap ranking analysis supported this interpretation for Northwest China (Fig. S8 and Table S14). The selected stage-level schemes had median top-three retention rates of 0.974 in station-bootstrap samples and 0.984 in time-block-bootstrap samples. The 95% Δ cMIEI intervals were positive for several nearest-competitor comparisons, especially in the land-surface and radiation stages, indicating stable separation from close alternatives in those cases.

Figure 3(g) presents the Taylor plot for the combined simulations of all 44 parameterization scheme combinations. Each color denotes a different physical process option, with the blue symbol representing the optimal scheme and the black symbol indicating the default scheme for the study area. The plot revealed a clear separation in the statistical outcomes associated with different physical processes. Following the systematic selection of parameterization schemes across five physical processes, the selected configuration showed a higher R and a lower RMSE than the default configuration. In the full-sample evaluation, the selected configuration performed better than the default configuration. The optimal scheme achieved a higher correlation coefficient ($R = 0.364$) compared to the default ($R = 0.295$), and a lower RMSE (1.277 versus 1.300), showing better evaluated agreement with the observed meteorological variables in the arid region of Northwest. A detailed analysis of individual physical processes indicated that microphysical, cumulus, and land surface schemes primarily contributed to reducing RMSE under the present evaluation metrics, though they exerted limited influence on the correlation coefficient. The planetary boundary layer process and the radiation process mainly affected the correlation coefficient and had a relatively weak influence on the root mean square error.

3.2 Comparison of proposed $MRE \times CV$ weighting and equal weighting methods in the arid region of Northwest

In addition to employing the proposed $MRE \times CV$ weighting method to determine the optimal parameterization scheme for the arid region of Northwest, this study also applies an equal-weighting approach, maintaining a consistent evaluation framework for both methods. Figure 4 presents the simulation results for various output variables obtained using the two weighting strategies.

Figure 4(a) displays the Taylor plots for the optimal parameterization schemes derived from the proposed $MRE \times CV$ weighting and equal-weighting methods, providing a visual comparison of their respective performances. The optimal scheme identified through the proposed $MRE \times CV$ weighting achieved a lower cMIEI value (1.165) compared to that derived from the equal-weighting method (1.184). More specifically, the proposed weighting approach yielded a higher correlation coefficient ($R = 0.364$) versus the equal-weighting method ($R = 0.323$).

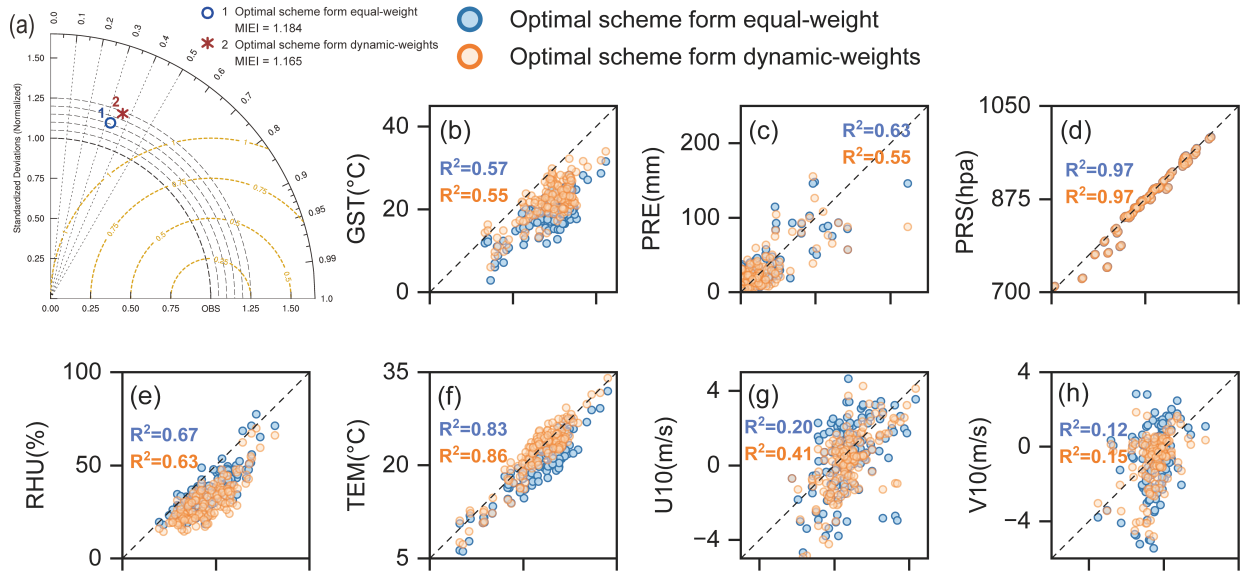


Figure 4. Comparison of simulation results obtained from the optimal parameterization schemes using the proposed $MRE \times CV$ weighting and equal weight methods. (a) Taylor plots of the optimal parameterization schemes obtained by different methods. (b-h) Scatter plots of the optimal parameterization schemes obtained via different methods for different meteorological elements. (b) Surface temperature, (c) precipitation, (d) Air pressure, (e) Relative humidity, (f) Air temperature, (g) Zonal wind, (h) Meridional wind. Panel (a) uses Taylor-diagram and cMIEI statistics calculated from daily station-level June-August values; panels (b-h) use monthly station summaries only for visual comparison, with monthly means for non-precipitation variables and monthly totals for precipitation.

Figure 4(b-h) shows scatter plots of the optimal parameterization schemes obtained via proposed $MRE \times CV$ weighting and equal weighting methods, which present a comparison of the monthly mean values (where precipitation is the monthly cumulative value) and the observed values of the two optimal parameterization schemes at different stations. These monthly summaries are used only for the scatter-plot visualization and are not the temporal aggregation used for the Taylor-diagram or cMIEI statistics. Among all the variables, air pressure performed the best. Both schemes showed extremely high fitting degrees and simulated the changing trend of air pressure relatively accurately. For the simulation results of surface temperature and relative humidity, the configuration selected by the proposed weighting scheme yielded a slightly lower R^2 compared to the equal weighting scheme. During the temperature simulation process, the configuration selected by the proposed weighting scheme demonstrated superior performance under high-temperature scenarios, achieving an R^2 of 0.86. In contrast to the air temperature results, the configuration selected by the proposed weighting scheme performed relatively poorly in the precipitation simulation, and R^2 decreased to 0.55. For the simulation of 10 m wind, neither scheme performed well in representing the zonal and meridional components. However, the R^2 values of the configuration selected by the proposed weighting scheme reached 0.41 and 0.15, which were higher than those of the equal weight scheme (0.20 and 0.12), indicating better evaluated

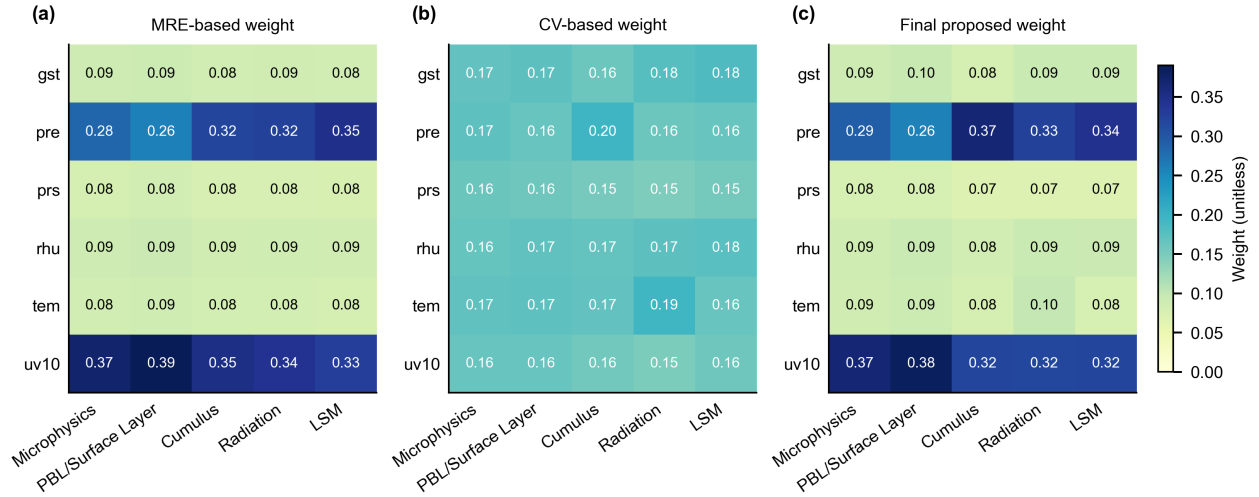


Figure 5. Weight distributions of the variables in the WRF simulations of the different physical processes in the coastal region of Southeast. (a) MRE-based weight, (b) CV-based weight, and (c) final proposed weight. Weights are unitless.

agreement for 10 m wind. Although the simulation performances of some variables have slightly decreased, on the whole, the parameterization combination selected by the proposed $\text{MRE} \times \text{CV}$ weights showed better overall agreement with observed meteorological variability in the arid region of Northwest China than the combination selected by the equal-weight method.

3.3 Optimal parameterization scheme combination for the coastal region of Southeast

Figure 5 shows the weight distributions of different variables across physical processes in the WRF simulations for the coastal region of Southeast, as determined by the proposed $\text{MRE} \times \text{CV}$ weighting method. Figure 5(a) represents the weights ω_{m1} calculated on the basis of the mean relative error ($\overline{\text{MRE}_m}$). The simulation weights for 10 m wind and precipitation remained the highest, with average weights of 0.36 and 0.31, respectively. Compared with those of the arid region of Northwest, the weights of the other variables were generally greater. The weight for the 10 m wind has decreased noticeably, indicating that the WRF model has reduced the disparity between the 10 m wind simulation and other variables in the coastal region of Southeast, suggesting a more balanced simulation capability across different variables in this region.

Figure 5(b) represents the weights ω_{m2} calculated on the basis of the coefficient of variation ($\text{CV}_{|\text{ME}|m}$) of the absolute error. There was substantial variability in the weights assigned to different variables throughout the simulation process, reflecting differences in the scheme-dependent response of variable errors within each physical-process evaluation stage. Precipitation exhibited the greatest variation in weight across physical processes, with a maximum value of 0.196 in the cumulus process, highlighting that precipitation errors responded strongly to alternative cumulus schemes. In contrast, air temperature and surface temperature showed stronger responses in the radiation process, with corresponding weight values of 0.194 and 0.177, respectively.

Figure 6(a) presents the median-relative cMIEI results for the evaluation of WRF parameterization schemes in the coastal region of Southeast. Analysis of the simulation performance across different variables revealed substantial variability in precipitation simulation across physical processes, with the microphysics process generally exhibiting the lowest performance. This outcome reflected the strong response of precipitation errors to alternative microphysics schemes. Within this category, the WDM6 scheme was identified as the optimal parameterization scheme and was subsequently incorporated into the following stages of evaluation. As a result, the parameterization schemes for the remaining physical processes demonstrated lower evaluated precipitation errors within the sequential screening framework. For 10 m wind, the planetary boundary layer and land surface processes exerted a more direct influence, yielding superior simulation results relative to other physical processes. Surface temperature, air temperature, and relative humidity exhibited substantial variation in cMIEI values throughout the evaluation. These variations were attributed to the complex influence of multiple physical processes and the dominant role played by previously selected optimal parameterization schemes.

The cMIEI values derived from the multivariate integrated evaluation showed a consistent decline as the assessment progresses, indicating that the selected configurations achieved progressively lower integrated evaluation scores at successive physical-process stages. Based on the multivariate evaluation, the optimal combination of parameterization schemes for the coastal region of Southeast included WDM6 (microphysics), Shin-Hong-MM5 (planetary boundary layer), Grell 3 (cumulus), GFDL-GFDL (radiation), and Noah LSM (land surface). As with the arid region of Northwest, the optimal combination does not guarantee superior simulation performance for all variables. For instance, in the cumulus process, the Grell 3 scheme performed exceptionally well in simulating precipitation and relative humidity but was slightly less effective for air temperature and 10 m wind. Nevertheless, due to the higher weight assigned to precipitation in the cumulus evaluation, Grell 3 remained the preferred scheme in the integrated analysis.

Figure 6(b-f) present Taylor diagrams for different physical process parameterization schemes. In the microphysical process, the optimal parameterization scheme, WDM6, had a relatively small RMSE (1.042), demonstrating good consistency between the simulated and observed values of anomalous fields. The correlation coefficients for all parameterization schemes within the microphysics process ranged from 0.42 to 0.47, exhibiting only minor variation, which suggested that the microphysics process exerted limited influence on the model's correlation performance. In contrast, within the planetary boundary layer process, the optimal scheme demonstrated a well-balanced simulation capability, with an RMSE of 1.066 and a correlation coefficient (R) of 0.488, both ranking among the highest across evaluated schemes. In the cumulus category, the Grell 3 scheme emerged as the optimal choice, yielding the lowest RMSE (1.032) and a relatively high correlation coefficient (0.478), while also effectively capturing the amplitude of anomalous field variations. For the radiation and land surface processes, the GFDL and Noah LSM schemes, respectively, achieved superior correlation coefficients compared to other schemes. Although their RMSE were relatively elevated, these schemes were nonetheless regarded as optimal due to their overall strong performance in the integrated evaluation.

For Southeast China, the selected schemes also generally remained in the leading group under resampling (Fig. S8 and Table S14). The median top-three retention rates were 0.822 in station-bootstrap samples and 0.944 in time-block-bootstrap samples.

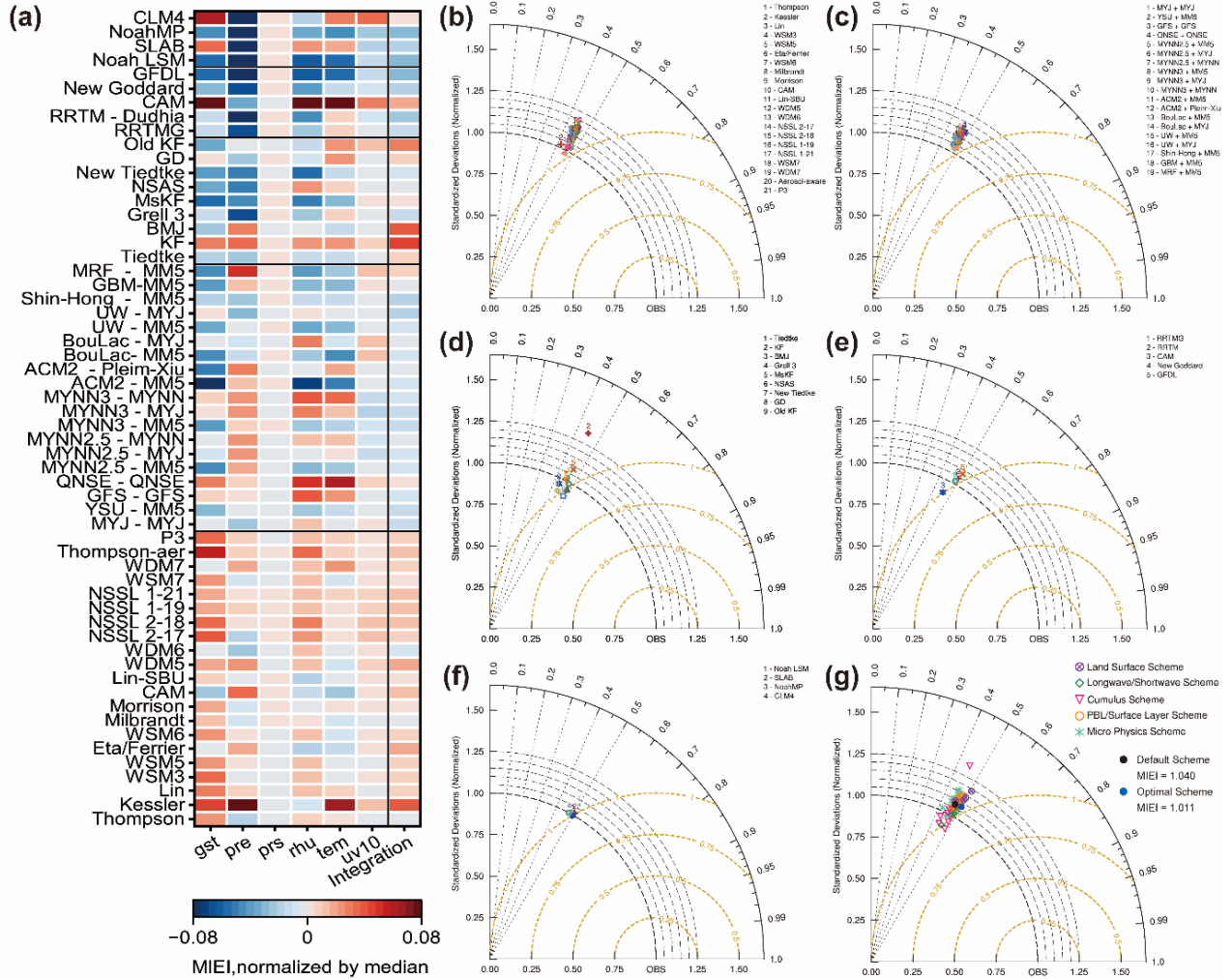


Figure 6. Evaluation results of parameterization schemes in the coastal region of Southeast. (a) Median-relative cMIEI, $(cMIEI - C_{med})/C_{med}$, where C_{med} is the regional reference median calculated separately for each variable column and the integrated column. This transformation is for visualization only; negative values indicate lower cMIEI. Colors saturate at ± 0.08 . (b-f) Taylor diagrams of different physical processes. (b) Microphysics, (c) Planetary boundary layer process, (d) Cumulus, (e) Radiation, (f) Land surface, and (g) Taylor diagrams of the candidate configurations pooled across all five physical-process stages. All cMIEI and Taylor-diagram statistics were calculated from daily station-level June-August values.

The lower retention in the PBL/surface-layer stage reflects the close performance of adjacent schemes, whereas the radiation and land-surface stages showed stronger top-three retention.

Figure 6(g) presents the Taylor diagram summarizing the combined simulation results of all 54 parameterization scheme combinations. Following the sequential selection of optimal schemes for the five physical processes, the selected configura-

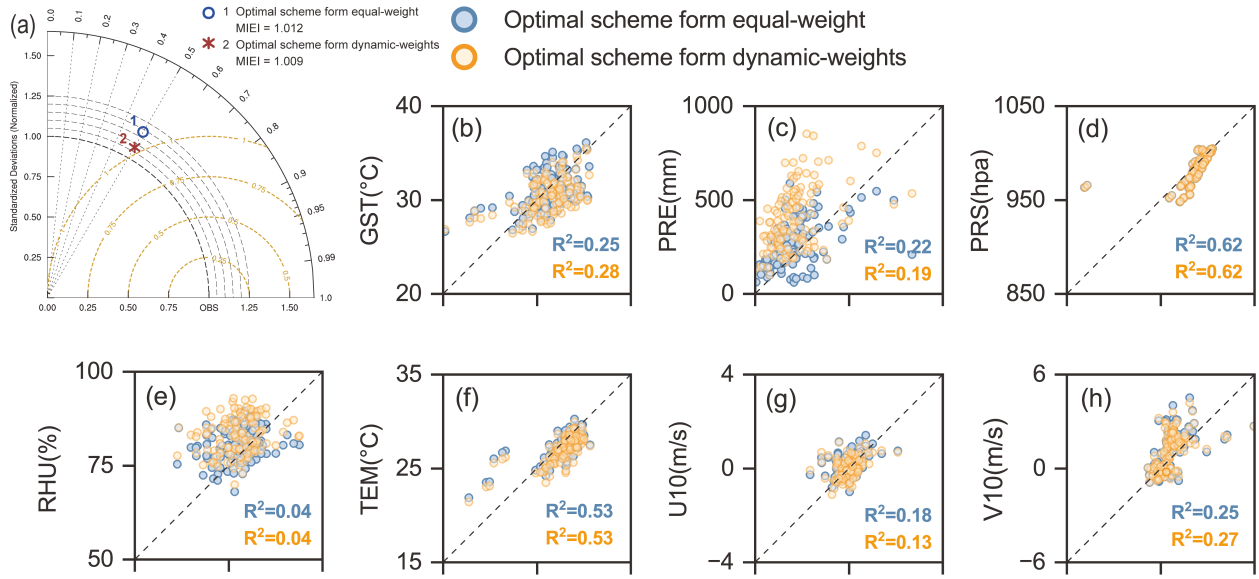


Figure 7. Comparison of simulation results obtained from the optimal parameterization schemes using the proposed $\text{MRE} \times \text{CV}$ weighting and equal weight methods. (a) Taylor plots of the optimal parameterization schemes obtained by different methods. (b-h) Scatter plots of the optimal parameterization schemes obtained via different methods for different meteorological elements. (b) Surface temperature, (c) precipitation, (d) Air pressure, (e) Relative humidity, (f) Air temperature, (g) Zonal wind, (h) Meridional wind. Panel (a) uses Taylor-diagram and cMIEI statistics calculated from daily station-level June-August values; panels (b-h) use monthly station summaries only for visual comparison, with monthly means for non-precipitation variables and monthly totals for precipitation.

tion achieved a lower RMSE and a higher R than the default configuration. Compared with the default scheme, the optimal configuration achieved a higher R value (0.502 versus 0.470) and a lower RMSE (1.039 versus 1.067), showing better evaluated agreement with the observed meteorological variables in the coastal region of Southeast. Independent analysis of each physical process revealed that microphysical, cumulus, and planetary boundary layer processes primarily contributed to reducing RMSE under the present evaluation metrics. In contrast, the radiation and land surface processes mainly influenced the correlation coefficient, with comparatively limited effects on the root mean square error.

3.4 Comparison of proposed $\text{MRE} \times \text{CV}$ weighting and equal weight methods in the coastal region of Southeast

Figure 7 compares the performance of the optimal parameterization schemes derived using the proposed $\text{MRE} \times \text{CV}$ weighting method and the equal-weighting method in simulating various output variables. Figure 7(a) displays the Taylor plots corresponding to each method. The optimal scheme obtained through the proposed $\text{MRE} \times \text{CV}$ weighting exhibited a lower cMIEI value (1.009) than that derived from the equal-weighting method (1.012), indicating stronger evaluated performance under the Taylor and cMIEI metrics.

Figure 7(b-h) presents scatter plots comparing the observed and simulated results for each output variable using the respective optimal parameterization schemes. As in Fig. 4, these scatter plots use monthly station summaries for visualization and do not define the temporal aggregation of the Taylor-diagram or cMIEI statistics. In the simulation of surface temperature, the configuration selected by the proposed weighting scheme outperformed the equal-weighting approach, with R^2 values of 0.28 and 0.25, respectively. For precipitation, the configuration selected by the proposed weighting scheme yielded an R^2 value of 0.19, slightly lower than that of the equal-weighting method. Air pressure emerged as the best-simulated variable under both schemes. The configuration selected by the proposed weighting scheme achieved an R^2 value of 0.62, capturing the variability in air pressure. In the case of relative humidity and air temperature, both schemes showed comparable correlation and similar performance. For wind variables, the configuration selected by the proposed weighting scheme showed differing performance between the zonal and meridional components. The R^2 values for zonal and meridional wind were 0.13 and 0.27, respectively, compared to 0.18 and 0.25 under the equal-weighting scheme. Overall, the relative performance of the parameterization combinations selected by the proposed $MRE \times CV$ weighting method and the equal-weighting approach varied across meteorological variables in the coastal region of Southeast China.

To further clarify how the proposed $MRE \times CV$ weights changed the scheme-selection pathway relative to equal weighting, Table S13 summarizes the highest-weight variables, Top-1 scheme changes, and variable-level contribution terms for rank-reversal cases. Across the ten region-stage evaluations, the Top-1 scheme changed in three stages: the Northwest PBL/surface-layer stage from MYJ + MYJ to YSU + MM5, the Northwest radiation stage from RRTMG to GFDL, and the Southeast cumulus stage from New Tiedtke to Grell 3. In all three cases, the pairwise cMIEI difference between the $MRE \times CV$ -weight winner and the equal-weight winner changed sign, from positive under equal weighting to negative under $MRE \times CV$ weighting ($D_{eq} = 0.0047, 0.0074, \text{ and } 0.0013$; $D_{dyn} = -0.0056, -0.0119, \text{ and } -0.0059$ for B1-B3, respectively). The variable-level contribution terms show that the higher UV10 weight promoted the selection of YSU + MM5 and GFDL in Northwest China, whereas the higher PRE weight promoted the selection of the Grell 3 cumulus scheme in Southeast China. Thus, the proposed $MRE \times CV$ weighting did not uniformly rescale all candidate schemes; it changed the integrated ranking when high-weight variables favored a different candidate scheme, preserving information on limiting variables and stage-specific scheme-dependent responses in the final screening pathway.

3.5 Comparison with established statistical weighting methods

To examine how the proposed method differs from established statistical weighting approaches in both weight structure and parameterization-scheme selection, we compared equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $MRE \times CV$ weighting. The comparison was conducted for the two representative screening years: 2007 in the arid region of Northwest China and 2009 in the coastal region of Southeast China. All methods used the same candidate schemes, observations, variables, and cMIEI evaluation framework. For each method, cMIEI, cVSC, cRMSL, R, and RMSE were computed from the same daily station-level June-August samples.

The complete stepwise selection pathways for Entropy, CV, and CRITIC weighting are reported in Table S17, including the
540 selected scheme and its stage-specific cMIEI at each of the five stages in both regions. These pathways document how the
successive choices produce the final configurations compared in Table 3.

Figure 8 compares the average variable-weight structures of the five weighting methods and the MRE-only baseline. Equal
weighting assigned the same value to each of the six evaluated variable groups. The CRITIC weights were also close to
a uniform distribution, with average weights ranging from 0.145 to 0.185 in Northwest China and from 0.150 to 0.197 in
545 Southeast China. Although CRITIC explicitly includes an inter-variable conflict/non-redundancy term (Table S1), its final
weights remained relatively close to equal weighting in this dataset.

Entropy weighting and coefficient-of-variation weighting showed more concentrated and mutually similar patterns because
both methods primarily respond to the dispersion of variable errors across candidate schemes. In Northwest China, both meth-
ods assigned high average weights to TEM, reaching 0.448 for entropy weighting and 0.356 for coefficient-of-variation weight-
550 ing, while assigning low weights to PRS and UV10. In Southeast China, both methods gave relatively high weights to GST,
TEM, and PRE. These similarities indicate that entropy weighting and coefficient-of-variation weighting identified broadly
similar discriminating variables in the present error matrices, although their numerical differences were sufficient to change
some scheme-ranking outcomes.

The proposed $\text{MRE} \times \text{CV}$ method produced a different weight structure. It consistently assigned larger weights to PRE and
555 UV10, with average weights of 0.240 and 0.431 in Northwest China and 0.317 and 0.342 in Southeast China, respectively.
This pattern reflects the purpose of the method: variables that are both difficult to simulate and responsive to alternative
parameterization choices within the evaluated physical-process group retain greater influence in the multivariable score. The
complete process-level weight matrices for the five methods are provided in Figs. S3-S4, and the sensitivity of the proposed
weights to the normalization function is shown in Fig. S5.

560 The MRE-only baseline reproduced the dominant weight pattern of the proposed method (Fig. 8). The average MRE-only
weights for PRE and UV10 were 0.239 and 0.445 in Northwest China and 0.307 and 0.357 in Southeast China, respectively.
Including CV introduced stage-dependent adjustments, with the largest absolute weight change of 0.048 occurring for PRE
in the Southeast cumulus stage. Some non-winning candidates exchanged adjacent ranks in the Southeast microphysics and
PBL/surface-layer stages. Table S11 in the Supporting Information reports the absolute weight changes for all six variable
565 groups and the schemes selected under both rules at each stage.

The resampling analyses showed that the dominant weight structure was stable to station and intra-seasonal temporal sam-
pling (Fig. S6 and Table S8). Across the ten region-stage combinations, the mean Spearman correlation with the full-sample
weights ranged from 0.947 to 1.000 for station bootstrap samples and from 0.944 to 1.000 for 14 d block bootstrap sam-
ples. The lowest lower-tail or monthly-subset correlation was 0.886. PRE and UV10 formed the set of the two highest-weight
570 variables in every station bootstrap, time-block bootstrap, and monthly subset. Their first and second positions occasionally
interchanged in several Southeast China stages, most clearly in the radiation stage, but the identity of the dominant two-variable
set was unchanged. For variable deletion, the recomputed retained-variable weights agreed with the theoretical renormalization
to a maximum absolute residual of 2.22×10^{-16} , and their ordering was preserved in all region-stage cases (Table S9). The

inter-variable dependence diagnostics showed that the strongest positive dependence occurred between GST and TEM. Their Pearson correlations were 0.930 in Northwest China and 0.847 in Southeast China for the observed daily station series, and 0.960 and 0.974, respectively, for the proposed scheme-level standardized-error matrices (Fig. S7). Removing GST while retaining TEM left the selected stage-level scheme unchanged in 8 of the 10 region-stage comparisons. The two changes occurred in the Southeast China PBL/surface-layer stage and land-surface stage (Table S12). Because the evaluated variable set differs between the original and no-GST analyses, the absolute cMIEI values are not interpreted as direct cross-set comparisons; the relevant evidence is the stability and physically interpretable sensitivity of the selected schemes.

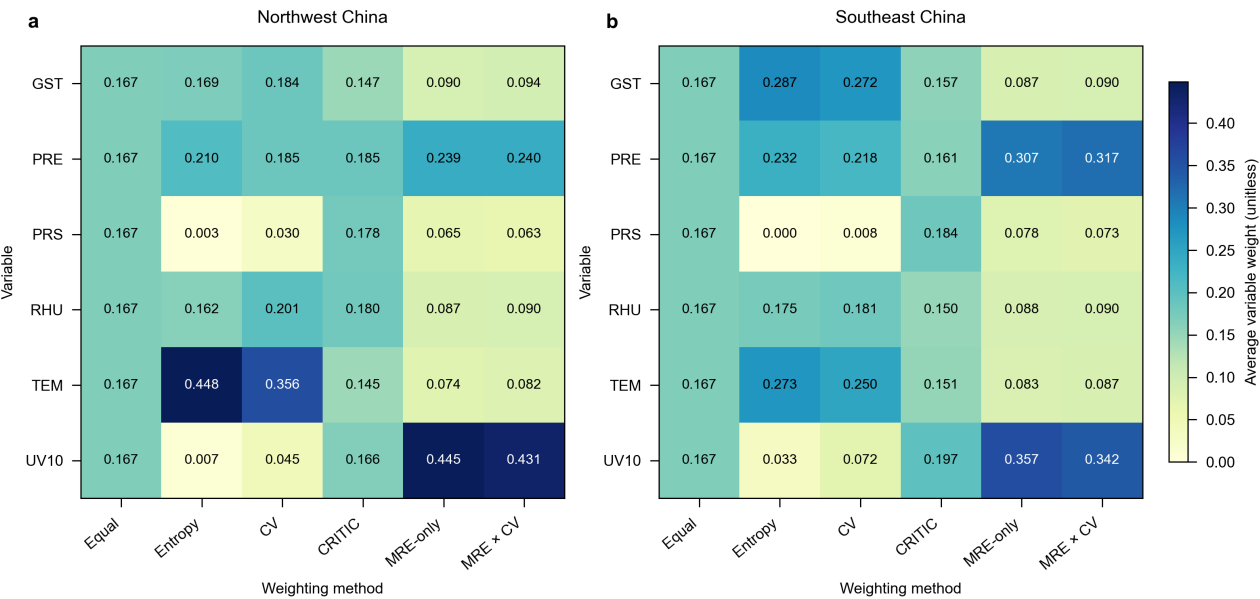


Figure 8. Comparison of average variable weights derived from equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, the MRE-only baseline, and the proposed MRE × CV weighting. Values represent averages across the five physical-process evaluation stages. The MRE-only and proposed weights were calculated from the same candidate set at each stage. (a) Arid region of Northwest China; (b) coastal region of Southeast China. Weights are unitless.

The final configurations in Tables 3 and 4 were evaluated using a common MRE × CV reference-weight vector within each region-year, calculated from the corresponding final-configuration error matrix. In Northwest China, the proposed MRE × CV method achieved the lowest cMIEI (1.166), compared with 1.186 for equal weighting and CRITIC, 1.198 for entropy weighting, and 1.207 for coefficient-of-variation weighting. In Southeast China, the proposed method also achieved the lowest cMIEI (1.006), slightly lower than coefficient-of-variation weighting (1.008), equal weighting and CRITIC (1.009), and entropy weighting (1.018). These results show that the proposed configuration achieved the most favorable integrated performance under the common MRE × CV evaluation objective.

Table 3: Performance comparison of the configurations selected by five weighting methods in the screening years. The selected configuration is identified by the weighting method used in the stepwise parameterization-screening workflow. Lower cMIEI indicates better integrated performance. The cMIEI, cVSC, and cRMSL values were calculated from daily station-level June-August values. Cross-evaluation results are provided in Table S16.

Region	Screening year	Weighting method	cMIEI	cVSC	cRMSL	Rank by cMIEI
Northwest China	2007	Equal weighting	1.186	0.321	1.167	2
		Proposed $MRE \times CV$	1.166	0.365	1.250	1
		Entropy weighting	1.198	0.365	1.428	4
		CV weighting	1.207	0.355	1.443	5
		CRITIC weighting	1.186	0.321	1.167	2
Southeast China	2009	Equal weighting	1.009	0.510	1.085	3
		Proposed $MRE \times CV$	1.006	0.507	1.091	1
		Entropy weighting	1.018	0.498	1.089	5
		CV weighting	1.008	0.506	1.016	2
		CRITIC weighting	1.009	0.510	1.085	3

We further tested whether the selected configurations were sensitive to the way in which the two proposed weight components were combined. The equally weighted additive alternative ($\alpha = 0.5$) selected the same stage-level schemes as the proposed multiplicative rule in all five Southeast China evaluation stages. In Northwest China, it selected the same schemes in four of the five stages, with a change only in the cumulus-stage scheme. The full stage-level results are provided in Table S10. Because the additive and multiplicative rules selected the same final configuration in Southeast China, no duplicate 2010 simulation was required for that region. In Northwest China, where the additive rule produced a different final configuration, the independent 2010 validation gave cMIEI = 1.312 for the additive-rule configuration, close to but slightly higher than the proposed multiplicative-rule configuration (cMIEI = 1.308). These results indicate that the main scheme-selection outcome was largely stable to the choice of combination rule, although the different weight allocations changed the cumulus-stage selection in Northwest China.

Variable-level performance differed between the two screening regions (Table S15). In Northwest 2007, the proposed configuration improved RMSE, MAE, absolute mean bias, and correlation for GST and TEM relative to Equal. TEM RMSE decreased from 3.686 to 2.704 °C, and GST RMSE decreased from 11.053 to 8.588 °C. The respective reductions were 0.98

and 2.47 °C, whereas RHU RMSE and MAE increased. Relative to the CV configuration in the same case, precipitation RMSE was slightly higher but MAE was lower.

In Southeast 2009, the proposed configuration reduced both RMSE and MAE for U10 and V10 relative to Equal. PRE and RHU showed increases in both error measures, with PRE RMSE increasing from 18.027 to 20.238 mm d⁻¹.

605 **3.6 Independent 2010 validation of the five weighting methods**

To examine the performance of the selected configurations outside the representative screening years, we further used 2010 as an independent extreme climate year for validation. Table 4 uses a common reference vector calculated for each 2010 regional evaluation, while Table S16 evaluates the five selected configurations under five common weighting objectives calculated separately for each region-year.

610 The 2010 validation results are summarized in Table 4. In Northwest China, the proposed $MRE \times CV$ method retained the lowest cMIEI (1.308), slightly lower than equal weighting and CRITIC (1.312), coefficient-of-variation weighting (1.322), and entropy weighting (1.325). In Southeast China, the proposed method again produced the lowest cMIEI (1.005), compared with 1.014 for equal weighting and CRITIC, 1.014 for entropy weighting, and 1.030 for coefficient-of-variation weighting. This independent extreme-year test provides additional evidence that the proposed weighting strategy can retain its integrated
615 performance advantage outside the screening year.

Table 4: Independent 2010 validation performance of the configurations selected by five weighting methods. Lower cMIEI indicates better integrated performance. The cMIEI, cVSC, and cRMSL values were calculated from daily station-level June-August values. Cross-evaluation under five common weighting objectives for each region-year is reported in Table S16.

Region	Validation year	Weighting method	cMIEI	cVSC	cRMSL	Rank by cMIEI
Northwest China	2010	Equal weighting	1.312	0.193	1.256	2
		Proposed $MRE \times CV$	1.308	0.184	1.152	1
		Entropy weighting	1.325	0.228	1.425	5
		CV weighting	1.322	0.223	1.412	4
		CRITIC weighting	1.312	0.193	1.256	2
Southeast China	2010	Equal weighting	1.014	0.507	1.077	3
		Proposed $MRE \times CV$	1.005	0.511	1.074	1
		Entropy weighting	1.014	0.504	1.075	2
		CV weighting	1.030	0.489	1.001	5
		CRITIC weighting	1.014	0.507	1.077	3

In Northwest 2010, the proposed configuration had lower precipitation and wind-component RMSE and MAE than Equal, and lower corresponding RMSE than Entropy and CV. In Southeast 2010, both wind components also had lower RMSE and MAE than Equal. These improvements were accompanied by increased errors in Northwest GST and TEM and Southeast PRE and RHU (Table S15), reflecting the multivariable performance trade-offs involved in configuration selection.

620 Cross-evaluation using weights recalculated for each region-year retained the proposed configuration as the lowest-cMIEI configuration under $MRE \times CV$ weighting in both 2010 regions (1.3082 in Northwest China and 1.0052 in Southeast China; Table S16). In Southeast 2010, it also ranked first under Equal and CRITIC evaluation weights. Other objectives favored alternative configurations, including Entropy in Northwest 2010 under Equal, Entropy, and CV weights, and the CV-selected configuration under CRITIC weights. These differences reflect the variable priorities encoded in each evaluation objective and

625 complement the unweighted variable-level comparisons in Table S15.

4 Discussion

4.1 Evaluation meaning and weight-allocation mechanism

The performance-based multivariable weighting framework proposed in this study introduces a two-component mechanism for assigning variable weights during WRF parameterization-scheme evaluation. It does not directly modify the WRF physical parameterizations; instead, it changes how different variables contribute to the integrated evaluation and thereby affects which scheme combinations are selected. Accordingly, lower cMIEI, RMSE, or higher R values should be interpreted as better evaluated performance of the selected configurations, not as direct improvement of WRF physics by the weighting method itself. As demonstrated in Figs. 2 and 5, the proposed framework identifies precipitation and 10 m wind as key constraining variables in two climatically distinct regions. This finding aligns with previous evaluation studies that have highlighted precipitation as a particularly challenging target variable (Yu et al., 2022b; Wang and Tan, 2023). This suggests that despite the different prevailing climate backgrounds, the WRF model may face common physical process challenges in accurately simulating precipitation associated with local strong convection and the near-surface wind field related to surface friction. In multivariable simulation studies, as the number of variables increases, those that are more difficult to simulate are often underrepresented in the evaluation process. The proposed framework addresses this issue by ensuring that critical variables with large simulation errors or strong scheme-dependent responses consistently receive appropriate weighting, thereby maintaining their influence and visibility throughout the model evaluation.

The comparison with entropy weighting, coefficient-of-variation weighting, and CRITIC further clarifies the evaluation meaning of the proposed weights. Entropy weighting and coefficient-of-variation weighting both emphasize dispersion among candidate schemes, which explains their similar weight patterns in the present experiments. CRITIC additionally considers inter-variable conflict, but the contrast and conflict components together produced only moderate departures from equal weighting in the present error matrices. By contrast, the proposed $\text{MRE} \times \text{CV}$ weighting is designed to address a specific scientific problem in WRF parameterization screening: key variables with large simulation errors may be diluted in a multivariable average, and different variables respond unevenly to alternative parameterization choices within each physical-process group. The MRE component represents variable-level simulation difficulty, and the absolute-error CV component represents the scheme-dependent response of variable errors. The resulting weights therefore jointly reflect simulation difficulty and the scheme-dependent response of variable errors within the evaluated physical-process group.

The multiplicative structure integrates these two diagnostic components through their product. For a fixed sum of the two component weights, their product increases as the difference between them decreases. Thus, among variables tied under the equally weighted additive rule, the product rule assigns a higher final weight to the variable with more balanced support from the two diagnostics. This property provides a rationale for using the product to combine simulation difficulty and scheme-dependent error response in parameterization screening.

The additive-rule sensitivity test assesses the practical consequences of this combination choice. The additive test produced the same selected configuration as the multiplicative rule in Southeast China and changed only the cumulus-stage choice in Northwest China. This pattern indicates that the main screening outcome was largely stable across the two combination rules,

660 while the Northwest China difference illustrates that their weight allocations can lead to different selections. In the independent 2010 Northwest China validation, the multiplicative-rule configuration had a slightly lower cMIEI than the additive-rule configuration (1.308 versus 1.312) under the same reference weights. These results support the use of the multiplicative combination in the present case studies. The dependence analysis also clarifies the boundary of the weighting framework. GST and TEM are physically related thermal variables, and treating highly correlated variables as independent information can duplicate their influence in a multivariable score. However, their high correlation does not imply identical physical-process information. GST is more directly connected to the surface energy balance, land-surface exchange, and radiative heating or cooling at the ground surface, whereas TEM represents the near-surface air-temperature response after surface forcing is coupled with surface-layer and boundary-layer mixing, radiative forcing, and transport. The two variables may therefore show different error responses to different physical-process options. This distinction is one reason for including the absolute-error CV component: it tests whether each variable's errors actually vary among the candidate schemes within the evaluated physical-process group, rather than assuming that physically related variables carry identical screening information. We did not apply formal decorrelation in the main evaluation because the variables are meteorologically defined outputs with distinct diagnostic meanings, and orthogonalized components would be less directly interpretable for WRF parameterization-scheme screening. Instead, we retained the physically defined variables and evaluated redundancy risk explicitly through correlation diagnostics, the CRITIC benchmark, and the no-GST sensitivity test. The largely unchanged stage-level selections indicate that the main screening outcome was not dominated by the GST-TEM dependence. The two changes in Southeast China were confined to the PBL/surface-layer and land-surface stages, which is consistent with the close connection of GST-related thermal information to near-surface turbulent exchange, land-atmosphere coupling, and surface energy fluxes. The unchanged cumulus and radiation choices further indicate that removing GST did not overturn the core precipitation/convection and radiation-related selection pathway in the humid coastal domain.

In multi-physics WRF simulations, variables differ substantially in simulation difficulty, as indicated by their mean relative errors. MRE dominates weight allocation in the present cases, whereas CV adds information about the relative dispersion of scheme-level mean absolute errors, allowing weighting to reflect stage-specific differences in error response (Table S11). As physical parameterizations improve, CV could have a more prominent relative influence if MRE differences across variables narrow while CV differences persist. Whether this adjustment affects scheme selection also depends on the performance differences among candidates.

The weights derived from the coefficient of variation of absolute errors did not exhibit strong regional differences, indicating that the scheme-dependent response of variable errors is closely related to the physical mechanisms represented by the candidate parameterization schemes. For instance, the pronounced response of precipitation errors within the microphysics and cumulus groups is consistent with previous studies showing that these processes strongly influence precipitation simulation (Tian et al., 2017; Liu et al., 2018). Similarly, the stronger response of surface temperature in the land surface group supports the value of the method in capturing physically meaningful links between parameterization choices and variable-level model performance (Liu et al., 2023).

4.2 Physical interpretation and practical meteorological relevance

695 The regional differences in the selected configurations provide a basis for interpreting how the weighting method responds to the dominant physical constraints in each climate setting. By giving greater influence to variables with larger relative errors and stronger scheme-dependent responses, the integrated evaluation links scheme selection to the variables that most constrain multivariable performance. The method evaluates the performance of existing parameterizations without changing their physical formulations. The resulting regional preferences can be understood in terms of how cloud, boundary-layer, and land-surface
700 processes influence the evaluated variables under contrasting moisture and thermal conditions.

In the arid region of Northwest China, limited water-vapor supply, complex terrain, strong land-atmosphere coupling, and large diurnal temperature variations make precipitation, near-surface wind, and temperature relevant to the evaluation of cloud microphysics, boundary-layer mixing, radiation, and surface energy-water exchanges. At the microphysics stage, P3 has the lowest precipitation cMIEI among the tested candidates, with a value of 1.090 compared with 1.192 for Thompson (Fig. 3a). P3
705 predicts bulk ice particle properties, including rimed ice mass and volume, allowing ice properties to evolve more continuously than in schemes with fixed ice categories (Morrison and Milbrandt, 2015). These evolving properties affect particle fall speeds and sedimentation, providing a flexible treatment of mixed-phase hydrometeors. Changes in the growth and fallout of ice particles can alter the amount and distribution of precipitation reaching the surface, offering a possible explanation for P3's lower precipitation cMIEI in this region, where intermittent precipitation is sensitive to cloud microphysical processes.

710 At the Northwest PBL/surface-layer stage, the selected YSU-MM5 combination has lower cMIEI for both UV10 and TEM than the MYJ-MYJ baseline (Fig. 3a). These variable-level gains contribute to its lower integrated cMIEI, although the integrated ranking reflects the balance across all evaluated variables. YSU uses a non-local K-profile approach with an explicit treatment of entrainment at the PBL top to represent turbulent exchange in the unstable mixed layer (Hong et al., 2006). This treatment redistributes heat and momentum through the boundary layer and controls exchange with the air above it, affecting
715 near-surface temperature and wind profiles. It therefore offers a possible explanation for the simultaneous wind and temperature gains of the selected combination. This interpretation is consistent with regional WRF studies showing sensitivity of heat and momentum fields to PBL choices and of arid-region summer climate to land-surface choices (Lu et al., 2021; Jia et al., 2023).

In the humid coastal region of Southeast China, the main simulation challenges are more closely related to monsoon moisture
720 transport, moist convection, cloud-precipitation processes, and the coastal boundary layer. WDM6 has the lowest precipitation cMIEI at the microphysics stage, with a value of 1.099 compared with 1.133 for P3 (Fig. 6a). WDM6 extends WSM6 by predicting cloud-water and rain number concentrations, together with cloud condensation nuclei, while retaining the single-moment treatment of ice species (Lim and Hong, 2010). The resulting double-moment treatment of cloud water and rain allows particle size distributions to respond to changes in both mass and number concentration. Rain number concentration affects
725 evaporation and the fraction of precipitation that reaches the surface (Lim and Hong, 2010). This offers a possible explanation for the lower precipitation cMIEI in the warm, moist Southeast environment.

At the Southeast PBL/surface-layer stage, the Shin–Hong–MM5 combination has lower TEM and GST cMIEI than the MYJ–MYJ baseline, with little change in precipitation cMIEI (Fig. 6a). The improvement in both temperature variables contributes to its integrated performance. Shin and Hong (2015) formulate scale-dependent vertical transport in the convective boundary layer, distinguishing local and non-local contributions to heat transport, while retaining YSU-type treatment of stable boundary-layer and free-atmospheric mixing. These transport terms alter the vertical distribution of heat and the thermal gradient between the surface and the atmosphere. In a coastal environment influenced by land-sea thermal contrasts and moist boundary-layer processes, the resulting changes in surface-atmosphere exchange may contribute to the lower TEM and GST cMIEI. The associated exchange of moisture and momentum also remains relevant to near-surface humidity and wind.

The selected Grell 3 scheme also has the lowest precipitation cMIEI at the Southeast cumulus stage: 1.053 compared with 1.068 for New Tiedtke (Fig. 6a). This result provides a direct variable-level link to its selection under the high precipitation weight at this stage. Grell 3 builds on the mass-flux and ensemble-based convection formulations described by Grell (1993) and Grell and Devenyi (2002), in which alternative closure and convective assumptions affect the redistribution of heat and moisture and the precipitation response. Closure assumptions determine how convective mass flux responds to the resolved atmospheric state; the associated heating and moistening tendencies influence the subsequent development of precipitation. The ensemble-based formulation samples different such responses, which may better match the range of convective conditions in the Southeast case and thus help explain its lower precipitation cMIEI. The importance of this response is consistent with previous findings that convection and microphysics choices affect Asian cloud and precipitation simulations and East Asian summer monsoon rainfall (Liu et al., 2018; Yang et al., 2015).

Sub-grid convection, radiative heating, and land-surface exchange complement these microphysical and boundary-layer processes. Tiedtke is retained at the Northwest cumulus stage on the basis of the integrated evaluation; its mass-flux treatment includes shallow and deeper convection and momentum transport, representing convective exchange that is not fully resolved by the model (Tiedtke, 1989; Zhang et al., 2011). At the radiation stage, GFDL has the lowest UV10 cMIEI in Northwest China and the lowest PRE and UV10 cMIEI in Southeast China (Figs. 3a and 6a). GFDL represents radiative transfer involving atmospheric gases and cloud effects, including the carbon dioxide treatment described by Fels and Schwarzkopf (1981). Differences in radiative heating can alter thermal structure and atmospheric stability, with subsequent effects on near-surface flow and precipitation development. Noah is subsequently retained in both regions on the basis of the integrated evaluation. Its representation of soil temperature, soil moisture, snow, and surface fluxes supplies a physically relevant link between land-surface conditions and near-surface meteorology (Chen and Dudhia, 2001). Together, radiative heating and the partitioning of available surface energy into sensible and latent heat influence boundary-layer development and its coupling to wind, temperature, and precipitation. These exchanges are relevant in both the arid inland and humid coastal domains, although their relative roles depend on water availability and the regional thermal background. The common selection of GFDL and Noah is consistent with this coupled influence within the tested configurations.

The selected configurations thus exhibit variable-level performance that is physically interpretable under the contrasting regional climate backgrounds. The mechanisms discussed above remain hypotheses for these responses, rather than direct

process attribution. Process-level diagnostics and multi-year validation would help establish how these mechanisms operate and whether the same combinations remain favorable across different circulation patterns and types of extreme events.

The variable-level results clarify the meteorological significance of the integrated rankings. Temperature improvements in Northwest 2007 concern the representation of near-surface thermal conditions, while wind improvements in Southeast 2009 and 2010 concern near-surface flow. These benefits coexist with higher errors in other variables, including Southeast precipitation and humidity and Northwest 2010 temperatures (Table S15). Their practical value therefore depends on which variables matter for the intended application and the magnitude of the associated error changes. Some RMSE increases coincide with lower MAE or higher correlation, so configuration performance needs to be assessed using these complementary measures. Considering error magnitudes, biases, and correlations together makes the balance between target-variable improvements and losses in other variables explicit.

4.3 Robustness and uncertainty boundaries of the screening framework

The representative-year screening and single independent extreme-year validation design used in this study was intended to balance statistical representativeness, computational feasibility, and the need for an out-of-sample test. The screening years were not chosen subjectively or only because their annual mean values were close to the climatological means. Instead, all candidate years from 1991 to 2020 were evaluated against the 30-year climatological reference using centered multivariable metrics. The years with the lowest cMIEI were selected for the two regions, and their variable-level metrics did not show pronounced anomalous deviations among the seven variables evaluated in the WRF assessment (Tables S2-S3 and Figs. S1-S2). Therefore, the selection was based on both the integrated cMIEI ranking and the absence of strong variable-level outliers, which reduces the risk that the parameterization screening is driven by an anomalous year. The 2010 experiment further provides a single independent extreme-year test of the selected schemes outside the screening period.

The representative-year screening and independent 2010 validation therefore provide two complementary tests: one under multivariable climatological representativeness and one under an extreme climate background outside the screening year. Broader multi-year optimization, leave-one-year-out cross-validation, or validation across additional climate-background years would further quantify the range of this robustness under different circulation regimes. In the present WRF physics-ensemble framework, such extensions would require rerunning the full parameterization screening across multiple years and regions. The present design therefore provides a reproducible and computationally tractable evaluation of the proposed weighting framework, while leaving broader multi-year generalization as a clear direction for future applications. The stability analysis specifically addresses uncertainty caused by station selection and intra-seasonal temporal sampling. The similar dominant weight structure obtained for the 2007 Northwest China and 2009 Southeast China cases provides supporting consistency across two distinct region-year settings, but region and year vary simultaneously and their effects cannot be separated. These comparisons should therefore not be interpreted as a controlled same-region cross-year weight-stability test.

The full-sample ranking first identifies the best-performing configuration for the evaluated station-time sample. The bootstrap analysis then examines how consistently that ranking is retained when the finite station and time samples are resampled.

The scheme-ranking analysis complements the weight-stability results by evaluating whether the selected stage-level
795 schemes remain competitive when the station and time samples are perturbed. Across the 20 region-stage-resampling comparisons, the selected schemes had a median top-three retention rate of 0.959. This result supports the use of the proposed framework to identify robust top-performing parameterization configurations in the evaluated station-time sample.

The WRF model offers many parameterization schemes that can be freely combined. However, this high degree of flexibility necessitates substantial computational resources to support the model's operation (Li et al., 2014). The stepwise simulation
800 method used in this study reduces the computational burden by evaluating physical-process options sequentially, while the interactions among WRF physical processes remain nonlinear and cannot be converted into truly independent additive effects. In multivariate evaluations, the intricate interactions between physical processes and variables therefore continue to introduce uncertainty (Zeng et al., 2016; Jia et al., 2023). To address the multivariable evaluation component of this problem, the proposed weighting framework adjusts variable weights during each stage of physical-process evaluation, ensuring that each process
805 emphasizes the variables most affected by simulation error and scheme-dependent response.

The theoretical implication of this design is that the search target changes from an exhaustive global optimization over all feasible WRF physics combinations to a sequence of conditional decisions. The scheme selected for an earlier process can modify the model state, the simulated error structure, and the performance landscape under which later processes are evaluated. For instance, as illustrated in Fig. 3(a), the selection of the YSU planetary boundary layer scheme is associated with lower
810 evaluated wind-speed errors in subsequent processes, indicating that earlier choices can propagate their influence into later stages. The resulting configuration should therefore be interpreted as the best-performing combination under the prescribed conditional screening framework, rather than as a mathematically guaranteed global optimum over all possible WRF physics combinations.

The prescribed order, from microphysics to planetary boundary layer, cumulus, radiation, and land surface processes, was
815 adopted as a practical and pre-defined screening sequence based on previous WRF sensitivity and physics-ensemble evaluation studies (Stergiou et al., 2017; Dai et al., 2021). In this sense, the order has both a procedural precedent and an empirical rationale: it evaluates cloud and precipitation microphysical processes, boundary-layer turbulent exchange, and convective processes before the radiation and land-surface choices are finalized. These earlier processes commonly exert strong direct or indirect controls on precipitation, near-surface wind, temperature, and humidity, and the present results likewise show substan-
820 tial multivariable differences associated with microphysics, cumulus, and planetary boundary layer choices. Nevertheless, the selected configuration remains conditional on the prescribed process order, because choices made at earlier stages determine the background configuration used to evaluate candidates at subsequent stages.

Within this conditional screening framework, the experiments evaluate the effects of physical parameterization choices while holding the remaining WRF settings fixed. The $\text{MRE} \times \text{CV}$ framework therefore provides a reproducible tool for
825 parameterization-scheme screening, rather than a complete decomposition of WRF simulation uncertainty.

4.4 Potential extension of the performance-based weighting framework to variable-level evaluation in multi-model ensembles

Beyond WRF parameterization-scheme evaluation, the performance-based weighting framework also suggests a possible extension for variable-level assessment in multi-model ensembles. The transferable component is not a direct attribution of structural uncertainty, but the use of observation-constrained error statistics to determine which variables should receive greater attention in a multivariable score. In the present WRF physics ensemble, the members share the same dynamical core, model configuration, domain setting, boundary conditions, and evaluation variables; therefore, the coefficient of variation of absolute errors can be interpreted mainly as the scheme-dependent response of each variable to alternative physical parameterization choices. In multi-model ensembles, such as CMIP-type ensembles, the same statistic would instead summarize variable-level inter-model spread, because differences among models integrate the effects of model structures, numerical schemes, spatial resolutions, physical process representations, tuning strategies, and model genealogy. This reinterpretation allows the performance-based weighting framework to be considered as a variable-level evaluation module for multi-model applications, while preserving the conceptual distinction between parameter perturbation within a single modelling framework and structural differences among distinct models.

Nevertheless, the proposed weighting idea can still provide value as a variable-level diagnostic component in future multi-model evaluation. Under this adapted interpretation, the mean relative error would identify variables with large systematic simulation errors across the ensemble, while the coefficient of variation of absolute errors would be interpreted as variable-level inter-model spread rather than within-process scheme-dependent response. Combining these two components would highlight variables that are both difficult to simulate and associated with large inter-model disagreement. Such variables, for example precipitation or near-surface wind in many regional climate applications, may otherwise be masked in an equal-weighted multivariable score by variables that are simulated more consistently. In this sense, the method could help preserve the visibility of high-error and high-spread variables during multi-variable model evaluation, while the final ensemble weighting would still require separate model-level treatment.

This distinction also defines the boundary of the method. The proposed variable weighting framework does not by itself attribute structural uncertainty to specific model components, nor does it solve model dependence, shared model genealogy, or redundancy among related model families. If applied to multi-model ensembles, it should be embedded in a broader hierarchical evaluation framework in which model outputs are first harmonized to common variables, spatial grids, temporal scales, and observational references, and then combined with model-level performance weights, model-independence weights, model-family grouping, or uncertainty decomposition approaches. Future work should therefore test this variable-level weighting module in multi-model ensembles and examine how it interacts with model-level weighting and structural-uncertainty attribution.

5 Conclusions

This study proposes a performance-based multivariable weighting framework for WRF parameterization-scheme evaluation. The method combines mean relative error and the coefficient of variation of absolute errors to give greater influence to variables

that are both difficult to simulate and responsive to alternative schemes within the evaluated physical-process group. Compared
860 with equal weighting, entropy weighting, coefficient-of-variation weighting, and CRITIC weighting, the proposed $MRE \times CV$
method produced the lowest integrated cMIEI in both representative screening years and in the independent 2010 extreme-year
validation.

Bootstrap analysis further showed that the selected stage-level schemes generally remained within the leading group under
station and time-block resampling. Across the 20 region-stage-resampling comparisons, the median top-three retention rate
865 was 0.959. These results indicate that the proposed weighting framework provides a transparent and reproducible basis for
selecting robust top-performing WRF parameterization configurations.

In the arid region of Northwest, the optimal parameterization scheme combination derived using the proposed $MRE \times CV$
weighting method includes P3, YSU, Tiedtke, GFDL, and Noah LSM. This configuration achieved lower integrated cMIEI
than the configurations selected by the benchmark weighting methods under the present evaluation framework. Similarly, in
870 the coastal region of Southeast China, the optimal combination comprised WDM6, Shin-Hong-MM5, Grell 3, GFDL-GFDL,
and Noah LSM, and it also achieved the lowest integrated cMIEI among the five weighting strategies. Compared with the
configuration selected by equal weighting, the configurations selected by the proposed method showed stronger evaluated
agreement with observed meteorological variability in several metrics. Nevertheless, remaining biases persisted, particularly
for precipitation and 10 m wind. These remaining biases indicate that the weighting method improves the evaluation and
875 selection of parameterization schemes, whereas further refinement of WRF physical parameterizations remains necessary for
representing complex weather systems and extreme events.

As parameterization schemes continue to evolve, multivariable evaluation methods need to account for both variable-level
simulation difficulty and scheme-dependent response within each physical-process evaluation stage. The proposed weighting
framework provides a reproducible way to incorporate these two aspects into WRF parameterization-scheme screening. Its
880 broader value therefore lies in improving the transparency and diagnostic relevance of multivariable model evaluation, rather
than replacing process-level model development or long-term validation.

Code and data availability. The Weather Research and Forecasting (WRF) model version 4.3.3 used in this study is freely available online
and can be downloaded from https://www2.mmm.ucar.edu/wrf/users/download/get_source.html (Skamarock et al., 2019). The code used
in this study to calculate the proposed $MRE \times CV$ weights is available at <https://doi.org/10.5281/zenodo.17414002> (Gou et al., 2025). The
885 Zenodo repository also provides the WRF model simulation results and the evaluation process. Any future updates or additional scripts will
be hosted there.

Author contributions. TG and SK conceptualized the study. TG and YD designed the methodology. TG developed the model code, performed
the simulations and drafted the manuscript. JN and SK supervised the work and contributed to the writing and editing of the paper.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

890 *Acknowledgements.* We thank Dr. Zhongfeng Xu from Institute of Atmospheric Physics, Chinese Academy of Sciences who provided guidance for MIEI methods.

Financial support. This research was supported by the National Key Research and Development Program of China (No. 2022YFD1900503). The computational work for this article was partially performed on resources of the National Supercomputing Centre in Tianjin.

References

- 895 Baker, N. C. and Taylor, P. C.: A framework for evaluating climate model performance metrics, *J. Clim.*, 29, 1773-1782, doi:10.1175/JCLI-D-15-0114.1, 2016.
- Borge, R., Alexandrov, V., Del Vas, J. J., Lumberras, J., and Rodríguez, E.: A comprehensive sensitivity analysis of the WRF model for air quality applications over the Iberian Peninsula, *Atmos. Environ.*, 42, 8560-8574, doi:10.1016/j.atmosenv.2008.08.032, 2008.
- Chen, F. and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I:
900 Model implementation and sensitivity, *Mon. Weather Rev.*, 129, 569-585, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2, 2001.
- Chen, F., Liu, C., Dudhia, J., and Chen, M.: A sensitivity study of high-resolution regional climate simulations to three land surface models over the western United States, *J. Geophys. Res. Atmos.*, 119, 7271-7291, doi:10.1002/2014JD021827, 2014.
- Dai, D., Chen, L., Ma, Z., and Xu, Z.: Evaluation of the WRF physics ensemble using a multivariable integrated evaluation approach over
905 the Haihe river basin in northern China, *Clim. Dyn.*, 57, 557-575, doi:10.1007/s00382-021-05723-x, 2021.
- Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., and Hoffman, F. M.: Taking climate model evaluation to the next level, *Nat. Clim. Change*, 9, 102-110, doi:10.1038/s41558-018-0355-y, 2019.
- Fels, S. B. and Schwarzkopf, M. D.: An efficient, accurate algorithm for calculating CO₂ 15 μ m band cooling rates, *J. Geophys. Res.*, 86, 1205-1232, doi:10.1029/JC086iC02p01205, 1981.
- 910 Fernández, J., Montávez, J. P., Sáenz, J., González-Rouco, J. F., and Zorita, E.: Sensitivity of the MM5 mesoscale model to physical parameterizations for regional climate studies: Annual cycle, *J. Geophys. Res. Atmos.*, 112, D04101, doi:10.1029/2005JD006649, 2007.
- Ferreira, J. A., Carvalho, A. C., Carvalheiro, L., Rocha, A., and Castanheira, J. M.: On the influence of physical parameterisations and domains configuration in the simulation of an extreme precipitation event, *Dyn. Atmos. Oceans*, 68, 35-55, doi:10.1016/j.dynatmoce.2014.08.001, 2014.
- 915 Fu, J., Kang, S., Zhang, L., Li, X., Gentile, P., and Niu, J.: Amplified warming induced by large-scale application of water-saving techniques, *Environ. Res. Lett.*, 17, 034018, doi:10.1088/1748-9326/ac4b52, 2022.
- Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, *J. Geophys. Res. Atmos.*, 113, D06104, doi:10.1029/2007JD008972, 2008.
- Grell, G. A.: Prognostic evaluation of assumptions used by cumulus parameterizations, *Mon. Weather Rev.*, 121, 764-787,
920 doi:10.1175/1520-0493(1993)121<0764:PEOAUB>2.0.CO;2, 1993.
- Grell, G. A. and Devenyi, D.: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques, *Geophys. Res. Lett.*, 29, 38-1-38-4, doi:10.1029/2002GL015311, 2002.
- Gou, T., Deng, Y., Niu, J., and Kang, S.: A performance-based multivariable weighting framework for WRF parameterization-scheme selection, *Zenodo* [code], <https://doi.org/10.5281/zenodo.17414002>, 2025.
- 925 Hausfather, Z., Drake, H. F., Abbott, T., and Schmidt, G. A.: Evaluating the performance of past climate model projections, *Geophys. Res. Lett.*, 47, e2019GL085378, doi:10.1029/2019GL085378, 2020.
- Hong, S.-Y., Noh, Y., and Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes, *Mon. Weather Rev.*, 134, 2318-2341, doi:10.1175/MWR3199.1, 2006.

- Jia, W., Zhang, X., Wang, H., Wang, Y., Wang, D., Zhong, J., Zhang, W., Zhang, L., Guo, L., and Lei, Y.: Comprehensive evaluation of typical planetary boundary layer (PBL) parameterization schemes in China–Part 1: Understanding expressiveness of schemes for different regions from the mechanism perspective, *Geosci. Model Dev.*, 16, 6635–6670, doi:10.5194/gmd-16-6635-2023, 2023.
- Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, *J. Clim.*, 23, 2739–2758, doi:10.1175/2009JCLI3361.1, 2010.
- Lambert, S. J. and Boer, G. J.: CMIP1 evaluation and intercomparison of coupled climate models, *Clim. Dyn.*, 17, 83–106, doi:10.1007/PL00013736, 2001.
- Li, L., Li, W., and Jin, J.: Improvements in WRF simulation skills of southeastern United States summer rainfall: physical parameterization and horizontal resolution, *Clim. Dyn.*, 43, 2077–2091, doi:10.1007/s00382-013-2031-2, 2014.
- Li, X., Wang, C., and Lan, J.: Role of the South China Sea in Southern China rainfall: Meridional moisture flux transport, *Clim. Dyn.*, 56, 2551–2568, doi:10.1007/s00382-020-05603-w, 2021.
- Lim, K.-S. S. and Hong, S.-Y.: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models, *Mon. Weather Rev.*, 138, 1587–1612, doi:10.1175/2009MWR2968.1, 2010.
- Liu, D., Yang, B., Zhang, Y., Qian, Y., Huang, A., Zhou, Y., and Zhang, L.: Combined impacts of convection and microphysics parameterizations on the simulations of precipitation and cloud properties over Asia, *Atmos. Res.*, 212, 172–185, doi:10.1016/j.atmosres.2018.05.017, 2018.
- Liu, L., Zhang, X., Han, C., and Ma, Y.: A performance evaluation of various physics schemes on the predictions of precipitation and temperature over the Tibet Autonomous Region of China, *Atmos. Res.*, 292, 106878, doi:10.1016/j.atmosres.2023.106878, 2023.
- Lu, S., Guo, W., Xue, Y., Huang, F., and Ge, J.: Simulation of summer climate over Central Asia shows high sensitivity to different land surface schemes in WRF, *Clim. Dyn.*, 57, 2249–2268, doi:10.1007/s00382-021-05876-9, 2021.
- Lv, M., Xu, Z., and Yang, Z.: Cloud resolving WRF simulations of precipitation and soil moisture over the central Tibetan Plateau: An assessment of various physics options, *Earth Space Sci.*, 7, e2019EA000865, doi:10.1029/2019EA000865, 2020.
- Mooney, P. A., Mulligan, F. J., and Fealy, R.: Evaluation of the sensitivity of the weather research and forecasting model to parameterization schemes for regional climates of Europe over the period 1990–95, *J. Clim.*, 26, 1002–1017, doi:10.1175/JCLI-D-11-00676.1, 2013.
- Morrison, H. and Milbrandt, J. A.: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests, *J. Atmos. Sci.*, 72, 287–311, doi:10.1175/JAS-D-14-0065.1, 2015.
- Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, *J. Geophys. Res. Atmos.*, 113, D14209, doi:10.1029/2007JD009334, 2008.
- Reichler, T. and Kim, J.: How well do coupled models simulate today’s climate?, *Bull. Am. Meteorol. Soc.*, 89, 303–312, doi:10.1175/BAMS-89-3-303, 2008.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, F.: Deep learning and process understanding for data-driven Earth system science, *Nature*, 566, 195–204, doi:10.1038/s41586-019-0912-1, 2019.
- Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., and de Mora, L.: Earth System model evaluation tool (ESMValTool) v2. 0–technical overview, *Geosci. Model Dev.*, 13, 1179–1199, doi:10.5194/gmd-13-1179-2020, 2020.
- Skamarock, W., Klemp, J., Dudhia, J., Gill, D. O., Liu, Z., Berner, A. J., Wang, A. W., Powers, A. J. G., Duda, A. M. G., Barker, A. D., and Huang, A. X.-Y.: A Description of the Advanced Research WRF Model Version 4.3, NCAR \[code\], <https://doi.org/10.5065/1dfh-6p97>, 2019.

- Shin, H. H. and Hong, S.-Y.: Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions, *Mon. Weather Rev.*, 143, 250-271, doi:10.1175/MWR-D-14-00116.1, 2015.
- 970 Stergiou, I., Tagaris, E., and Sotiropoulou, R.-E. P.: Sensitivity assessment of WRF parameterizations over Europe, *Proceedings*, 1, 119, doi:10.3390/ecas2017-04138, 2017.
- Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res. Atmos.*, 106, 7183-7192, doi:10.1029/2000JD900719, 2001.
- Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, *Mon. Weather Rev.*, 117, 1779-1800, doi:10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2, 1989.
- 975 Tian, J., Liu, J., Yan, D., Li, C., and Yu, F.: Numerical rainfall simulation with different spatial and temporal evenness by using a WRF multiphysics ensemble, *Nat. Hazards Earth Syst. Sci.*, 17, 563-579, doi:10.5194/nhess-17-563-2017, 2017.
- Wang, C., Qian, Y., Duan, Q., Huang, M., Yang, Z., Berg, L. K., Gustafson Jr, W. I., Feng, Z., Liu, J., and Quan, J.: Quantifying physical parameterization uncertainties associated with land-atmosphere interactions in the WRF model over Amazon, *Atmos. Res.*, 262, 105761, doi:10.1016/j.atmosres.2021.105761, 2021.
- 980 Wang, X. and Tan, Z.: On the combination of physical parameterization schemes for tropical cyclone track and intensity forecasts in the context of uncertainty, *J. Adv. Model. Earth Syst.*, 15, e2022MS003381, doi:10.1029/2022MS003381, 2023.
- Xu, Z., Han, Y., and Fu, C.: Multivariable integrated evaluation of model performance with the vector field evaluation diagram, *Geosci. Model Dev.*, 10, 3805-3820, doi:10.5194/gmd-10-3805-2017, 2017.
- Xu, Z., Hou, Z., Han, Y., and Guo, W.: A diagram for evaluating multiple aspects of model performance in simulating vector fields, *Geosci. Model Dev.*, 9, 4365-4380, doi:10.5194/gmd-9-4365-2016, 2016.
- 985 Yang, B., Zhang, Y., Qian, Y., Huang, A., and Yan, H.: Calibration of a convective parameterization scheme in the WRF model and its impact on the simulation of East Asian summer monsoon precipitation, *Clim. Dyn.*, 44, 1661-1684, doi:10.1007/s00382-014-2118-4, 2015.
- Yu, E., Ma, J., and Sun, J.: Developing a climate prediction system over Southwest China using the 8-km weather research and forecasting (WRF) model: system design, model calibration, and performance evaluation, *Weather Forecast.*, 37, 1703-1719, doi:10.1175/WAF-D-21-0188.1, 2022a.
- 990 Yu, E., Bai, R., Chen, X., and Shao, L.: Impact of physical parameterizations on wind simulation with WRF V3.9.1.1 under stable conditions at planetary boundary layer gray-zone resolution: a case study over the coastal regions of North China, *Geosci. Model Dev.*, 15, 8111-8134, doi:10.5194/gmd-15-8111-2022, 2022b.
- Zeng, X., Wang, B., Zhang, Y., Zheng, Y., Wang, N., Wang, M., Yi, X., Chen, C., Zhou, Z., and Liu, H.: Effects of land surface schemes on WRF-simulated geopotential heights over China in summer 2003, *J. Hydrometeorol.*, 17, 829-851, doi:10.1175/JHM-D-14-0239.1, 2016.
- 995 Zhang, C., Wang, Y., and Hamilton, K.: Improved representation of boundary layer clouds over the Southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme, *Mon. Weather Rev.*, 139, 3489-3513, doi:10.1175/MWR-D-10-05091.1, 2011.
- Zhang, M., Xu, Z., Han, Y., and Guo, W.: An improved multivariable integrated evaluation method and tool (MVIETool) v1.0 for multimodel intercomparison, *Geosci. Model Dev.*, 14, 3079-3094, doi:10.5194/gmd-14-3079-2021, 2021.
- 1000 Zhang, X., Zhao, T., Xu, H., Liu, W., Wang, J., Chen, X., and Liu, L.: GLC_FCS30D: The first global 30-m land-cover dynamic monitoring product with a fine classification system from 1985 to 2022 using dense time-series Landsat imagery and continuous change-detection method, *Earth Syst. Sci. Data Discuss.*, 16, 1353-1381, doi:10.5194/essd-16-1353-2024, 2024.