

Responses to Reviewers' Comments for Manuscript egusphere-2025-5362

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

Addressed Comments for Publication to
Geoscientific Model Development
by

Tianyu Gou, Yaoyang Deng, Jun Niu, Shaozhong Kang

Dear Editor and Referees,

We sincerely thank the editor for coordinating the peer-review process and for giving us the opportunity to revise our manuscript. We are also grateful to the three referees for their careful evaluation, constructive comments, and detailed suggestions. These comments helped us substantially improve the framing, methodological transparency, robustness assessment, and physical interpretation of the study.

In the revised manuscript, we have made several major changes. First, following the referees' comments on the use of the terms "dynamic" and "objective", we changed the manuscript title to "A performance-based multivariable weighting framework for WRF parameterization-scheme selection". This title change is intended to better reflect the revised framing of the method: the proposed $\text{MRE} \times \text{CV}$ weighting is derived from model-observation performance statistics and is used as a transparent evaluation-stage weighting framework for WRF parameterization-scheme selection.

Second, we strengthened the methodological description of the weighting framework, including the Softmax transformation, the final normalization of the combined weights, the rationale for the multiplicative formulation, and the relationship between the proposed method and benchmark weighting approaches, including equal weighting, entropy weighting, coefficient-of-variation weighting, and CRITIC weighting.

Third, we added a series of new analyses in the revised manuscript and Supporting Information to further strengthen the methodological transparency and robustness evidence of the study. These include benchmark-method comparisons, sensitivity tests for key weighting choices, sampling and variable-set stability analyses, inter-variable dependence diagnostics, and ranking-pathway analysis.

Finally, we clarified the representative-year selection procedure, expanded the interpretation of the independent 2010 extreme-year validation, and added explanatory text to the Discussion and Conclusions to articulate the scope of the proposed framework more clearly. We also added physical interpretations of the selected parameterization-scheme combinations and expanded the Discussion on the potential use of the framework as a variable-level evaluation module in multi-model ensembles, including the reinterpretation of the MRE and CV components under structural differences among models.

Our point-by-point responses to the referees' comments are provided below.

In the following responses, our replies are shown in black regular font, [whereas the corresponding changes in the revised manuscript and Supporting Information are shown in blue regular font.](#)

Sincerely,
Prof. Shaozhong Kang
(on behalf of all co-authors)
State Key Laboratory of Efficient Utilization of Agricultural Water Resources
Beijing 100083, China
Email: kangsz@cau.edu.cn

Response to Reviewer 1

General Comments

The authors present a novel framework for selecting physical parameterization schemes within the Weather Research and Forecasting (WRF) model. The study proposes a "dynamic weighting" approach that assigns importance to output variables based on two statistical criteria: simulation accuracy (Mean Relative Error) and sensitivity to physical processes (Coefficient of Variation of the absolute error). This method is applied to two climatically distinct regions in China and compared against a traditional equal-weighting strategy. The authors further validate their approach using an independent extreme weather year.

The manuscript is generally well-structured and addresses a relevant issue in regional climate modeling: the subjectivity involved in evaluating multivariate model performance. The proposed tool offers a potential pathway toward more reliable, data-driven parameterization selection and uncertainty reduction. However, the methodology requires further justification regarding the mathematical formulation of the weights, the sequential selection methodology, and the physical interpretation of the results. I recommend a moderate revision before possible publication in GMD, addressing these points detailed below.

Response to General Comments:

We sincerely thank the reviewer for the positive and constructive overall assessment of our manuscript. We appreciate the reviewer's recognition that the study addresses an important issue in WRF parameterization-scheme selection and multivariable model evaluation. We also thank the reviewer for clearly identifying the main aspects that required further justification, especially the mathematical formulation of the weights, the ordered scheme-selection strategy, and the physical interpretation of the selected configurations.

In response to these comments, we revised the manuscript substantially. We clarified the role of the proposed $\text{MRE} \times \text{CV}$ weighting framework and avoided overstating it as a fully objective or simulation-online dynamic method. We strengthened the mathematical description of the weighting procedure, including the Softmax normalization, the final normalization of the combined weights, and the rationale for the multiplicative formulation. We also added sensitivity analyses to test the effects of normalization choices and alternative component-combination rules.

We further clarified the representative-year selection procedure, the use of the independent 2010 validation year, and the limitations of single-year optimization. In addition, we expanded the explanation of the ordered screening strategy, explicitly discussed its path dependence, and clarified that the selected configuration should be interpreted as the best-performing combination under the prescribed conditional screening framework rather than as a mathematically guaran-

teed global optimum. Finally, we added a more detailed physical interpretation of the selected parameterization schemes in relation to the contrasting climatic characteristics of Northwest and Southeast China, and we revised the Discussion and Conclusions to define the scope and limitations of the proposed framework more carefully.

The above summarizes our overall response to the reviewer's general comments. Our detailed point-by-point responses to the reviewer's specific comments are provided below.

In the point-by-point responses below, our responses are shown in regular black font, whereas the corresponding revisions made in the manuscript are shown in blue regular font.

Comment 1.1

The authors employ the Softmax function to normalize statistical metrics into weights, a mathematical choice that inherently tends to polarize values. While the authors state this is to "amplify the impact of key limiting variables," this exponential scaling risks excessively penalizing variables with only marginally higher errors or sensitivities, potentially reducing a multivariate evaluation to a quasi-univariate one dominated by a single variable like precipitation. To establish the scientific robustness of this approach, it is essential to provide a justification for selecting Softmax over linear normalization techniques and to include a sensitivity analysis to ensure the method does not artificially skew the evaluation.

Response 1.1:

We thank the reviewer for raising this important issue. In the revised manuscript, we clarified the rationale for using Softmax and added an explicit sensitivity analysis to quantify whether the resulting weights were overly concentrated or dominated by a single variable.

The proposed evaluation framework was designed to prevent variables with large simulation errors or strong scheme-dependent responses from being diluted in a multivariable average. To test the robustness of this choice, we compared the original Softmax setting with linear normalization and temperature-scaled Softmax formulations. The original method corresponds to $\tau = 1$; smaller τ values produce more concentrated weights, whereas larger τ values produce smoother weights. The results show that the original Softmax setting provides an intermediate and balanced weighting behavior. Under $\tau = 1$, the mean maximum variable weight across the two regions and five physical processes was 0.394, and the effective number of variables was 3.85. This indicates that multiple variables still contributed substantially to the evaluation.

By contrast, linear normalization did not reduce the concentration of weights in this dataset. It produced a larger mean maximum weight of 0.538, with the maximum reaching 0.870 in some process-level cases, and the effective number of variables decreased to as low as 1.31. This result occurs because linear normalization directly transfers the magnitude contrast in the original statistical scores into the final weights; when precipitation has a much larger score than

the other variables, it can receive a very high weight. A more concentrated Softmax setting ($\tau = 0.5$) also reduced the effective number of variables, whereas smoother settings ($\tau = 2$ and $\tau = 5$) distributed weights more evenly but weakened the intended emphasis on limiting variables.

These results support the original Softmax setting as a balanced choice between emphasizing limiting variables and preserving the multivariable nature of the evaluation. We revised Section 2.3.3 to clarify the mathematical role of Softmax and the final normalization of the combined weights, and added the normalization sensitivity analysis to the Supporting Information as Fig. S5.

Corresponding revisions:

Section 2.3.3 Performance-based multivariable weighting framework (page 12, lines 252-254):

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. 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.

Supporting Information, Fig. S5:

Sensitivity of the proposed variable weights to the normalization function. (a) Mean variable weights under linear normalization and temperature-scaled Softmax transformations. The original Softmax setting corresponds to $\tau = 1$. (b) Weight-concentration diagnostics, including the maximum variable weight and the effective number of variables. Error bars indicate the range across the two study regions and five physical processes.

Supporting material cited in this response: Figs. S5

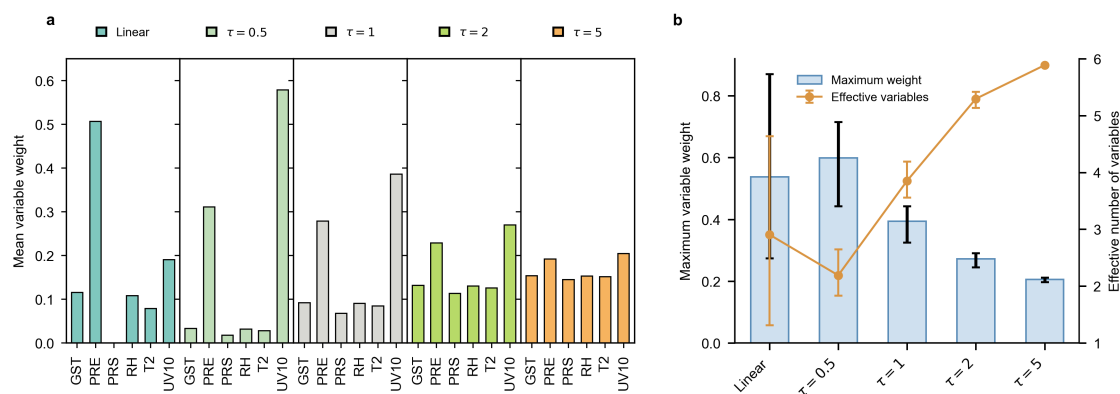


Figure S5. Sensitivity of the proposed variable weights to the normalization function. (a) Mean variable weights under linear normalization and temperature-scaled Softmax transformations. The original Softmax setting corresponds to $\tau=1$. (b) Weight-concentration diagnostics, including the maximum variable weight and the effective number of variables. Error bars indicate the range across the two study regions and five physical processes.

Comment 1.2

The study selects specific individual years to represent the climatological mean based on 30-year averages. However, relying on a single year for parameterization optimization may inadvertently overfit the model to specific synoptic patterns unique to that year, rather than capturing the region's broader climatological variability. The authors need to provide detailed assessment methods and quantitative analysis to prove that the selected year is statistically representative across all assessment variables, rather than merely representing the average state. Furthermore, a discussion regarding the potential biases introduced by single-year optimization versus multi-year validation is necessary to contextualize the universality of the selected schemes.

Response 1.2:

We thank the reviewer for highlighting the need to clarify how the screening years were selected and how the possible dependence on a single year should be interpreted. In the revised manuscript, we further clarified that the selected years are not interpreted as covering all synoptic conditions within a 30-year period, but as screening years closest to the 30-year climatological state under a long-term observation-based procedure. We also used an independent extreme year to evaluate the out-of-sample performance of the selected configurations.

In the revised manuscript, we expanded the description of the representative-year selection method in Section 2.2. Each candidate year from 1991 to 2020 was compared with the cor-

responding 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. We also clarified that the lowest cMIEI alone does not prove that a year is representative for every WRF evaluation variable. Therefore, the selection was interpreted together with the variable-level metrics: the selected years did not show pronounced outlying deviations among the seven variables evaluated in the WRF assessment and, overall, ranked among the candidate years closest to the long-term climatological state.

We also added Figs. S1-S2 to the Supplementary Information to show the full multivariable assessment for all candidate years from 1991 to 2020. These results support the selection of 2007 for the arid region of Northwest China and 2009 for the coastal region of Southeast China because these years had the lowest cMIEI values in their respective regions and did not show strong variable-level anomalies. These figures complement Tables S2-S3 by showing both the integrated metrics and the variable-level metrics used in the selection procedure.

We further revised the Discussion to clarify the potential bias and scope of the single-year optimization design. We now state that a single representative year cannot fully cover all interannual variability or synoptic-scale circulation patterns. A strict multi-year optimization or leave-one-year-out cross-validation design would provide a more complete test of cross-year stability. The present study uses a representative-year optimization plus independent extreme-year validation design; therefore, the selected configurations should be interpreted as better-performing parameterization combinations within the current screening and validation framework, rather than as globally optimal configurations for all possible years.

Corresponding revisions:

[Section 2.2 Model configuration and datasets \(page 5, lines 119-123\):](#)

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).

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 32, lines 728-735):

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.

Supporting Information, Figs. S1-S2:

Figure S1. Multivariable assessment of each candidate year from 1991 to 2020 against the 30-year climatological mean in the arid region of Northwest China using centered MVIE statistics. Lower cMIEI indicates closer agreement with the 30-year climatological reference. The year 2007 had the lowest cMIEI and did not show strong anomalous deviations in the variable-level metrics; it was therefore selected as the representative screening year for this region. Figure S2. Multivariable assessment of each candidate year from 1991 to 2020 against the 30-year climatological mean in the coastal region of Southeast China using centered MVIE statistics. Lower cMIEI indicates closer agreement with the 30-year climatological reference. The year 2009 had the lowest cMIEI and did not show strong anomalous deviations in the variable-level metrics; it was therefore selected as the representative screening year for this region.

Supporting material cited in this response: Figs. S1–S2 and Tables S2–S3

METRICS		1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
cRMSD	evp	0.407	0.306	0.220	0.290	0.194	0.348	0.270	0.298	0.262	0.368	0.359	0.381	0.430	0.295	0.278	0.175	0.209	0.313	0.342	0.299	0.313	0.246	0.527	0.351	0.812	0.597	0.489	0.674	0.727	0.908
	gst	0.109	0.135	0.150	0.112	0.088	0.154	0.120	0.143	0.182	0.212	0.151	0.141	0.163	0.123	0.097	0.148	0.104	0.199	0.116	0.173	0.241	0.126	0.182	0.070	0.151	0.206	0.100	0.278	0.181	0.181
	pre	0.988	0.528	1.215	0.836	1.062	0.830	0.607	0.798	0.354	0.639	1.210	0.207	0.906	0.740	1.050	0.739	0.300	0.649	0.922	0.981	0.824	0.557	0.716	0.745	0.862	0.590	0.826	0.546	1.315	0.531
	prs	0.219	0.354	0.654	0.539	0.276	0.392	0.523	0.329	0.579	0.151	0.366	0.581	0.155	0.331	0.493	0.405	0.452	0.135	0.310	0.599	0.894	0.147	0.155	0.297	0.314	0.163	0.313	0.627	0.361	0.311
	rhu	0.632	0.628	0.633	0.448	0.674	0.423	0.679	0.450	0.296	0.528	0.827	0.316	0.552	0.450	0.371	0.308	0.187	0.458	0.543	0.239	0.358	0.409	0.361	0.543	0.540	0.197	0.753	0.243	0.550	0.356
	ssd	0.448	0.626	0.786	0.842	0.418	0.620	0.701	0.920	0.377	0.982	1.154	0.812	0.771	0.571	0.564	0.149	0.295	0.471	0.553	0.398	0.361	0.285	0.217	0.615	0.454	0.756	0.832	0.323	0.773	0.955
	tem	0.135	0.178	0.177	0.132	0.119	0.201	0.125	0.173	0.235	0.246	0.168	0.182	0.215	0.143	0.112	0.179	0.158	0.257	0.150	0.233	0.361	0.139	0.281	0.073	0.181	0.288	0.145	0.380	0.225	0.227
	win	0.302	0.367	0.181	0.260	0.350	0.218	0.313	0.152	0.252	0.372	0.539	0.295	0.365	0.182	0.164	0.211	0.290	0.262	0.194	0.372	0.102	0.109	0.094	0.293	0.367	0.276	0.330	0.120	0.384	0.605
cRMSVD		0.501	0.325	0.587	0.425	0.512	0.435	0.341	0.418	0.261	0.397	0.615	0.244	0.487	0.382	0.503	0.370	0.201	0.377	0.464	0.494	0.463	0.298	0.431	0.389	0.536	0.417	0.461	0.446	0.694	0.488
SD_std		0.290	0.155	0.223	0.138	0.228	0.261	0.131	0.163	0.090	0.155	0.220	0.160	0.352	0.177	0.288	0.112	0.137	0.186	0.228	0.134	0.176	0.180	0.276	0.226	0.317	0.222	0.313	0.280	0.508	0.336
cMIEI		0.579	0.369	0.618	0.444	0.514	0.500	0.365	0.414	0.243	0.403	0.626	0.294	0.536	0.420	0.541	0.389	0.235	0.427	0.504	0.514	0.457	0.350	0.525	0.457	0.590	0.479	0.540	0.543	0.755	0.596
SD	evp	1.353	1.274	1.063	1.255	1.186	1.241	1.252	1.170	1.227	1.353	1.273	0.714	0.607	0.879	0.795	0.835	0.804	0.925	0.748	0.726	0.816	0.886	0.538	0.673	0.549	0.544	0.542	0.413	0.507	1.103
	gst	1.076	1.011	0.898	1.090	1.037	1.072	0.984	1.020	1.122	1.173	1.059	0.929	1.038	1.029	0.971	0.974	1.047	0.975	0.922	0.990	1.156	1.104	0.820	0.967	0.970	0.975	1.084	0.786	0.933	1.051
	pre	0.392	1.234	1.506	0.839	1.588	0.659	0.980	1.452	1.084	0.978	1.347	1.093	1.851	1.264	1.756	1.041	1.265	0.719	1.523	1.018	1.316	0.569	1.548	0.803	1.706	1.084	1.312	1.392	2.267	1.149
	prs	1.186	1.200	1.383	1.324	1.021	0.752	1.228	0.948	0.978	1.034	0.821	0.987	1.047	1.184	1.373	0.832	1.192	1.110	1.001	1.129	1.331	1.019	0.930	1.011	1.095	0.977	0.901	1.006	0.984	1.218
	rhu	1.274	1.339	1.005	1.051	1.507	1.016	0.864	0.994	0.995	1.244	1.528	0.740	0.781	0.962	0.988	1.055	1.155	1.393	1.239	1.203	1.143	1.103	0.800	1.477	1.052	1.037	1.015	0.862	0.932	1.119
	ssd	1.336	1.387	1.383	1.129	0.897	1.563	1.051	1.332	1.177	1.383	1.444	1.210	0.846	1.462	1.014	1.080	0.972	0.994	0.984	1.105	1.211	1.169	0.869	1.123	1.008	1.416	1.701	1.157	0.545	1.109
	tem	1.094	1.033	0.889	1.083	1.065	1.114	0.943	1.078	1.177	1.169	1.023	0.890	1.062	1.009	0.969	0.985	1.092	0.913	0.904	1.086	1.273	1.097	0.726	0.975	0.962	0.976	1.136	0.698	0.937	1.057
	win	1.226	0.930	1.108	1.202	1.250	1.049	1.152	1.135	0.997	0.940	1.235	1.054	1.254	1.035	0.988	1.181	0.981	0.865	0.894	1.065	0.905	0.959	1.008	1.145	0.705	0.932	0.954	1.029	1.001	0.913
cRMSL		1.049	1.121	1.089	1.074	1.220	1.049	1.023	1.168	1.141	1.162	1.173	0.929	1.211	1.059	1.156	0.983	1.075	0.911	1.055	0.999	1.174	0.976	0.968	0.923	1.109	0.952	1.090	0.898	1.286	0.988
CORR	evp	0.985	0.993	0.979	0.992	0.999	0.975	0.996	0.974	0.993	0.996	0.979	0.956	0.975	0.959	0.978	0.998	0.997	0.950	0.964	0.990	0.960	0.973	0.941	0.988	0.585	0.864	0.972	0.868	0.719	0.903
	gst	0.997	0.991	0.993	0.998	0.997	0.991	0.993	0.990	0.992	0.994	0.991	0.992	0.988	0.993	0.996	0.989	0.996	0.980	0.996	0.985	0.985	0.998	0.999	0.998	0.989	0.979	0.999	0.980	0.985	0.986
	pre	0.226	0.909	0.595	0.599	0.754	0.565	0.813	0.851	0.945	0.791	0.502	0.984	0.974	0.811	0.849	0.739	0.992	0.762	0.811	0.527	0.780	0.891	0.931	0.678	0.928	0.842	0.777	0.948	0.973	0.887
	prs	0.994	0.964	0.898	0.930	0.963	0.939	0.910	0.944	0.829	0.990	0.938	0.829	0.990	0.968	0.962	0.918	0.930	0.997	0.952	0.848	0.741	0.990	0.990	0.956	0.959	0.987	0.951	0.805	0.934	0.980
	rhu	0.873	0.896	0.801	0.906	0.934	0.912	0.744	0.898	0.956	0.912	0.867	0.978	0.836	0.896	0.930	0.957	0.995	0.980	0.904	0.993	0.953	0.929	0.943	0.977	0.863	0.982	0.721	0.977	0.840	0.950
	ssd	0.967	0.912	0.830	0.693	0.909	0.978	0.767	0.724	0.953	0.704	0.607	0.745	0.663	0.961	0.844	0.993	0.956	0.889	0.845	0.933	0.965	0.977	0.983	0.838	0.898	0.859	0.941	0.966	0.642	0.594
	tem	0.996	0.985	0.989	0.995	0.995	0.988	0.993	0.989	0.990	0.986	0.986	0.988	0.980	0.990	0.994	0.984	0.992	0.968	0.993	0.978	0.978	0.996	0.998	0.998	0.984	0.958	0.999	0.962	0.975	0.977
	win	0.984	0.930	0.991	0.989	0.976	0.979	0.967	0.998	0.968	0.928	0.904	0.960	0.973	0.985	0.986	0.995	0.957	0.971	0.985	0.937	0.999	0.995	0.996	0.972	0.966	0.961	0.944	0.993	0.926	0.803
cVSC		0.882	0.959	0.845	0.918	0.912	0.911	0.944	0.937	0.979	0.943	0.852	0.971	0.920	0.933	0.901	0.931	0.984	0.926	0.899	0.878	0.922	0.955	0.905	0.921	0.876	0.910	0.906	0.895	0.844	0.879

Figure S1. Multivariable assessment of each candidate year from 1991 to 2020 against the 30-year climatological mean in the arid region of Northwest China using centered MVIE statistics. Lower cMIEI indicates closer agreement with the 30-year climatological reference. The year 2007 had the lowest cMIEI and did not show strong anomalous deviations in the variable-level metrics; it was therefore selected as the representative screening year for this region. This climatological comparison was used for representative-year selection and was not the temporal aggregation used for WRF scheme-evaluation metrics.

METRICS		1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
cRMSD	gst	0.096	0.185	0.097	0.191	0.103	0.190	0.179	0.166	0.175	0.140	0.123	0.217	0.158	0.089	0.165	0.119	0.192	0.204	0.260	0.195	0.233	0.145	0.181	0.170	0.135	0.143	0.151	0.184	0.150	0.206
	pre	0.467	0.800	0.489	0.688	0.631	0.501	0.562	0.681	0.605	0.725	0.529	0.657	0.448	0.605	0.641	0.535	0.590	0.779	0.370	0.778	0.532	0.540	0.621	0.534	0.924	0.715	0.496	0.586	0.586	0.546
	prs	0.139	0.222	0.120	0.243	0.142	0.197	0.198	0.185	0.263	0.129	0.186	0.145	0.138	0.168	0.215	0.212	0.236	0.195	0.263	0.318	0.269	0.186	0.097	0.149	0.110	0.287	0.128	0.194	0.138	0.218
	rhu	1.103	1.500	0.705	1.064	1.114	0.798	0.960	0.985	1.530	0.996	0.965	1.052	0.931	0.755	1.062	0.362	1.069	0.968	1.098	1.051	1.221	1.082	1.000	1.255	1.284	1.240	1.024	1.497	1.478	0.953
	ssd	0.543	0.716	0.503	0.697	0.533	0.427	0.860	0.746	0.742	0.677	0.500	0.739	0.680	0.548	0.518	0.374	0.749	0.542	0.530	0.856	0.590	0.680	0.561	0.809	0.884	0.776	0.710	0.858	0.988	0.663
	tem	0.085	0.181	0.106	0.188	0.087	0.176	0.144	0.140	0.143	0.132	0.114	0.175	0.123	0.098	0.174	0.121	0.186	0.219	0.261	0.169	0.244	0.140	0.162	0.143	0.107	0.139	0.139	0.146	0.088	0.201
	win	2.237	1.740	1.612	2.353	1.194	2.601	1.735	1.425	1.801	2.080	1.397	2.353	1.550	0.898	1.057	2.009	1.567	1.651	1.515	2.048	2.089	1.134	1.249	1.441	1.428	1.231	1.893	1.804	1.043	2.323
cRMSVD		0.420	0.653	0.414	0.587	0.504	0.433	0.542	0.579	0.542	0.596	0.438	0.578	0.435	0.486	0.513	0.426	0.533	0.612	0.382	0.665	0.482	0.481	0.509	0.514	0.751	0.604	0.472	0.559	0.582	0.500
SD_std		0.562	0.412	0.340	0.611	0.233	0.689	0.353	0.387	0.491	0.455	0.314	0.457	0.355	0.187	0.186	0.429	0.279	0.258	0.248	0.531	0.418	0.282	0.243	0.352	0.262	0.131	0.353	0.350	0.394	0.503
cMIEI		0.769	0.799	0.561	0.856	0.549	0.894	0.647	0.705	0.773	0.758	0.531	0.742	0.619	0.535	0.499	0.614	0.615	0.600	0.485	0.838	0.696	0.572	0.566	0.659	0.775	0.615	0.607	0.682	0.692	0.750
SD	gst	1.013	1.051	1.016	0.951	1.063	1.041	0.878	1.045	0.922	1.023	0.970	0.901	1.061	1.001	1.112	0.980	0.999	1.087	1.042	1.002	1.135	1.099	1.009	1.082	0.937	1.047	0.978	0.994	0.948	0.943
	pre	0.743	1.247	1.315	1.401	1.212	1.149	1.163	1.257	1.208	1.033	1.364	1.112	1.018	0.935	1.423	1.294	1.158	1.503	0.918	1.403	0.921	0.939	1.102	1.299	1.175	1.010	1.245	0.962	1.323	0.991
	prs	1.025	1.056	1.052	1.049	1.049	0.963	0.989	0.890	1.077	0.973	1.021	1.030	1.022	0.999	1.076	1.009	1.023	1.083	1.008	0.801	1.102	1.032	0.993	1.041	1.024	1.067	0.960	1.074	1.057	0.987
	rhu	1.410	2.180	1.563	1.152	1.767	1.551	0.996	1.626	2.035	0.842	1.374	0.719	1.630	1.216	1.612	1.111	1.702	1.425	1.411	1.508	1.328	1.050	1.733	1.929	1.084	1.137	1.536	1.081	2.153	1.346
	ssd	1.377	1.537	1.085	1.185	1.211	1.079	0.778	1.484	0.826	1.311	1.060	0.749	1.389	1.075	1.326	1.076	1.264	1.024	1.063	1.489	1.205	1.477	1.243	1.401	1.061	1.189	1.217	1.015	1.560	0.998
	tem	0.998	1.027	1.028	0.964	1.059	1.058	0.914	1.014	0.947	1.012	0.978	0.901	1.036	0.982	1.122	0.975	0.972	1.113	1.039	0.972	1.146	1.099	1.004	1.064	0.941	1.040	0.974	1.014	0.960	0.966
	win	2.579	1.798	1.992	2.814	1.246	3.043	1.913	2.079	2.073	2.279	1.898	2.156	1.995	1.511	1.405	2.266	1.625	1.754	1.657	2.519	2.290	1.785	1.212	1.886	1.752	1.418	1.999	2.018	1.290	2.425
cRMSL		0.970	1.239	1.191	1.239	1.165	1.117	1.027	1.229	1.075	1.081	1.200	0.997	1.100	0.980	1.312	1.165	1.125	1.310	0.990	1.297	1.055	1.097	1.097	1.249	1.084	1.053	1.160	0.990	1.257	0.985
CORR	gst	0.996	0.985	0.996	0.982	0.997	0.983	0.990	0.988	0.987	0.991	0.993	0.979	0.990	0.996	0.993	0.993	0.982	0.984	0.968	0.981	0.984	0.995	0.984	0.990	0.992	0.991	0.989	0.983	0.990	0.975
	pre	0.898	0.768	0.947	0.889	0.855	0.900	0.876	0.842	0.867	0.746	0.946	0.811	0.902	0.807	0.918	0.923	0.861	0.882	0.929	0.842	0.850	0.847	0.830	0.925	0.650	0.747	0.925	0.822	0.910	0.850
	prs	0.991	0.978	0.994	0.973	0.992	0.981	0.980	0.988	0.971	0.992	0.983	0.990	0.991	0.986	0.981	0.978	0.973	0.986	0.966	0.961	0.972	0.984	0.995	0.990	0.994	0.964	0.992	0.985	0.993	0.976
	rhu	0.628	0.803	0.943	0.519	0.815	0.893	0.538	0.822	0.688	0.426	0.712	0.285	0.856	0.784	0.767	0.946	0.809	0.734	0.633	0.719	0.480	0.444	0.867	0.816	0.243	0.331	0.752	0.034	0.802	0.707
	ssd	0.944	0.927	0.887	0.809	0.901	0.918	0.556	0.891	0.685	0.862	0.884	0.677	0.888	0.863	0.939	0.938	0.806	0.857	0.870	0.834	0.873	0.921	0.897	0.824	0.633	0.762	0.812	0.638	0.788	0.780
	tem	0.996	0.984	0.995	0.982	0.998	0.987	0.993	0.990	0.991	0.992	0.994	0.988	0.993	0.995	0.993	0.993	0.983	0.984	0.968	0.986	0.983	0.995	0.987	0.992	0.996	0.992	0.990	0.990	0.997	0.980
	win	0.513	0.335	0.595	0.601	0.452	0.574	0.431	0.792	0.496	0.410	0.699	0.026	0.646	0.820	0.661	0.463	0.365	0.385	0.438	0.625	0.410	0.812	0.375	0.658	0.579	0.528	0.353	0.450	0.611	0.305
cVSC		0.910	0.851	0.943	0.884	0.903	0.922	0.857	0.885	0.866	0.839	0.937	0.833	0.919	0.880	0.937	0.934	0.881	0.894	0.926	0.864	0.891	0.899	0.886	0.919	0.743	0.828	0.915	0.842	0.892	0.874

Figure S2. Multivariable assessment of each candidate year from 1991 to 2020 against the 30-year climatological mean in the coastal region of Southeast China using centered MVIE statistics. Lower cMIEI indicates closer agreement with the 30-year climatological reference. The year 2009 had the lowest cMIEI and did not show strong anomalous deviations in the variable-level metrics; it was therefore selected as the representative screening year for this region. This climatological comparison was used for representative-year selection and was not the temporal aggregation used for WRF scheme-evaluation metrics.

Table S2 The comparison results of daily meteorological elements in the arid region of Northwest each year with the 30-year average climate values

Year	cRMSE	R	Standard Deviation Ratio	MIEI
1991	0.50	0.88	1.05	0.58
1992	0.32	0.96	1.12	0.37
1993	0.59	0.85	1.09	0.62
1994	0.42	0.92	1.07	0.44
1995	0.51	0.91	1.22	0.51
1996	0.44	0.91	1.05	0.50
1997	0.34	0.94	1.02	0.37
1998	0.42	0.94	1.17	0.41
1999	0.26	0.98	1.14	0.24
2000	0.40	0.94	1.16	0.40
2001	0.61	0.85	1.17	0.63
2002	0.24	0.97	0.93	0.29
2003	0.49	0.92	1.21	0.54
2004	0.38	0.93	1.06	0.42
2005	0.50	0.90	1.16	0.54
2006	0.37	0.93	0.98	0.39
2007	0.20	0.98	1.08	0.23
2008	0.38	0.93	0.91	0.43
2009	0.46	0.90	1.06	0.50
2010	0.49	0.88	1.00	0.51
2011	0.46	0.92	1.17	0.46
2012	0.30	0.95	0.98	0.35
2013	0.43	0.90	0.97	0.53
2014	0.39	0.92	0.92	0.46
2015	0.54	0.88	1.11	0.59
2016	0.42	0.91	0.95	0.48
2017	0.46	0.91	1.09	0.54
2018	0.45	0.89	0.90	0.54
2019	0.69	0.84	1.29	0.76
2020	0.49	0.88	0.99	0.60

Table S3 The comparison results of daily meteorological elements in the coastal region of Southeast each year with the 30-year average climate values

Year	cRMSE	R	Standard Deviation Ratio	MIEI
1991	0.42	0.91	0.97	0.77
1992	0.65	0.85	1.24	0.80
1993	0.41	0.94	1.19	0.56
1994	0.59	0.88	1.24	0.86
1995	0.50	0.90	1.17	0.55
1996	0.43	0.92	1.12	0.89
1997	0.54	0.86	1.03	0.65
1998	0.58	0.89	1.23	0.70
1999	0.54	0.87	1.08	0.77
2000	0.60	0.84	1.08	0.76
2001	0.44	0.94	1.20	0.53
2002	0.58	0.83	1.00	0.74
2003	0.43	0.92	1.10	0.62
2004	0.49	0.88	0.98	0.53
2005	0.51	0.94	1.31	0.50
2006	0.43	0.93	1.17	0.61
2007	0.53	0.88	1.12	0.62
2008	0.61	0.89	1.31	0.60
2009	0.38	0.93	0.99	0.49
2010	0.67	0.86	1.30	0.84
2011	0.48	0.89	1.05	0.70
2012	0.48	0.90	1.10	0.57
2013	0.51	0.89	1.10	0.57
2014	0.51	0.92	1.25	0.66
2015	0.75	0.74	1.08	0.78
2016	0.60	0.83	1.05	0.61
2017	0.47	0.92	1.16	0.61
2018	0.56	0.84	0.99	0.68
2019	0.58	0.89	1.26	0.69
2020	0.50	0.87	0.99	0.75

Comment 1.3

The authors employed an ordered selection strategy. As discussed in the relevant section (lines 571-579), this approach inevitably introduces path dependence. The optimal choice for the initial process inevitably constrains the performance landscape for subsequent choices, suggesting that the final combination likely represents a local optimum dependent on the arbitrary order of evaluation rather than a global optimum. The authors should critically discuss the theoretical implications of this limitation and, if possible, justify the specific order chosen.

Response 1.3:

We thank the reviewer for raising this important point. We agree that an ordered selection strategy can introduce path dependence, because the scheme selected for an earlier physical process defines the background configuration under which subsequent physical processes are evaluated. In the original manuscript, we had briefly noted this issue, but we agree that the theoretical implications of this limitation and the rationale for the prescribed order needed to be discussed more critically.

In the revised manuscript, we clarified in Section 2.2 that the sequence used in this study, namely microphysics, planetary boundary layer, cumulus, radiation, and land surface processes, was adopted as a pre-specified conditional screening strategy to reduce the computational burden associated with the high-dimensional WRF physics-configuration space. 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, radiation, and land surface scheme groups, and the WRF physics-ensemble evaluation of Dai et al. (2021). We also added that the order is consistent with empirical evidence from WRF ensemble evaluations and the present results, which show that microphysics, planetary boundary layer, and cumulus schemes often exert strong influences on summer near-surface meteorological variables, including precipitation, wind speed, temperature, and humidity.

We further expanded the Discussion to explain the theoretical implication of this strategy. The stepwise design changes the search target from an exhaustive global optimization over all feasible WRF physics combinations to a sequence of conditional decisions. Earlier choices can modify the model state, the simulated error structure, and the performance landscape for subsequent choices. Therefore, the selected configuration should be interpreted as the best-performing combination under the prescribed conditional screening framework, rather than as a mathematically guaranteed global optimum. We also clarified that the proposed weighting framework improves the multivariable evaluation within each step, but does not remove the path dependence inherent in sequential screening. Future applications could quantify this source of uncertainty by comparing alternative process orders, applying multi-start stepwise screening, or

using global optimization and surrogate-model approaches when computational resources allow.

Corresponding revisions:

Section 2.2 Model configuration and datasets (page 6, lines 140-147):

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 microphysics, 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, radiation, 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 precipitation, wind speed, temperature, and humidity, which are key variables in the present multivariable evaluation. Therefore, 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.

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 33, lines 763-766):

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 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 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 pro-

cesses, 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 substantial multivariable differences associated with microphysics, cumulus, and planetary boundary layer choices. Nevertheless, the order is not theoretically unique. The proposed weighting framework helps reduce evaluation bias within each step by emphasizing variables with larger errors or stronger scheme-dependent responses, but it does not eliminate the path dependence of the stepwise search itself.

Comment 1.4

While the manuscript presents comprehensive statistical evidence, it lacks sufficient analysis of the underlying physical mechanisms driving these results. The authors are encouraged to discuss the physical mechanisms of the optimal parameterization schemes by linking the method's characteristics with the regional climatic features. This will enhance the theoretical depth of the paper.

Response 1.4:

Thank you for this important suggestion. We agree that the original manuscript relied too heavily on statistical performance metrics and did not sufficiently explain why the selected parameterization schemes are physically meaningful under the two regional climate backgrounds.

In the revised manuscript, we added a new subsection, Sect. 4.2, “Physical interpretation and practical meteorological relevance”. This subsection explicitly links the proposed performance-based weighting framework to regional physical interpretation. Specifically, we now explain that the proposed weights emphasize variables with large relative errors and strong scheme-dependent responses, and that these high-weight variables can be interpreted as key constraints on regional model performance. We then connect these variables and their scheme-dependent responses to the dominant climate characteristics of the arid Northwest and humid coastal Southeast domains. To avoid overinterpretation, we also clarify that this discussion provides mechanism-informed physical consistency, not direct process-attribution evidence.

We supported this revision with technical and regional WRF literature. For example, we now discuss the physical characteristics of the selected P3, YSU, Tiedtke, GFDL, Noah LSM, WDM6, Shin-Hong, and Grell 3 schemes, and relate these characteristics to cloud microphysics, boundary-layer mixing, convective transport, radiative transfer, and land-surface fluxes in the two regions.

Corresponding revisions:

Section 4.2 Physical interpretation and practical meteorological relevance (page 30, lines 665-672):

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. The weighting framework itself does not directly diagnose process budgets or alter WRF physical parameterizations. Because the method gives greater influence to variables with larger relative errors and stronger scheme-dependent responses within the evaluated physical-process group, the final scheme combinations are not determined only by the lowest aggregate score; they also reflect which variables and processes most strongly limit the multivariable simulation in each region. The selected schemes should therefore be interpreted as statistically preferred configurations whose physical characteristics are consistent with the present regional climate backgrounds, rather than as universally optimal schemes or as proof of a single causal mechanism.

Section 4.2 Physical interpretation and practical meteorological relevance (page 31, lines 673-676):

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 particularly sensitive to cloud microphysics, boundary-layer mixing, radiation, and surface energy-water exchanges. The selection of the P3 microphysics scheme is physically plausible because P3 predicts bulk ice particle properties and explicitly represents the mass and volume of rimed ice, allowing a more flexible treatment of mixed-phase hydrometeors than schemes with fixed ice categories (Morrison and Milbrandt, 2015). This is relevant for precipitation simulation over arid and topographically complex regions, where precipitation events are intermittent and strongly affected by cloud microphysical uncertainty. The selection of the YSU PBL scheme is also consistent with the importance of near-surface wind and temperature in this region, because YSU uses a non-local K-profile approach with an explicit entrainment layer to represent turbulent exchange in the unstable mixed layer (Hong et al., 2006). The Tiedtke cumulus scheme, a mass-flux scheme with shallow convection and momentum transport, provides a physically meaningful representation of sub-grid convective transport under conditions where convection is not fully resolved (Tiedtke, 1989; Zhang et al., 2011). Together with the GFDL radiation scheme, which accounts for radiative transfer with carbon dioxide, ozone, and cloud microphysical effects (Fels and Schwarzkopf, 1981), and the Noah LSM, which represents soil temperature, soil moisture, snow, and surface fluxes in the land-atmosphere system (Chen and Dudhia, 2001), this combination is consistent with the dominant controls on wind, temperature, and intermittent precipitation in the arid Northwest domain. This interpretation is also supported by previous WRF studies showing that land surface choices can strongly affect summer climate simulations in arid Central Asia and that PBL schemes in China differ substantially in their representation of heat and momentum fields (Lu et al., 2021; Jia et al., 2023). In the

humid coastal region of Southeast China, the dominant simulation challenges are more closely related to monsoon moisture transport, moist convection, cloud-precipitation processes, and the coastal boundary layer. The selection of WDM6 is physically interpretable because this scheme extends WSM6 by including double-moment treatment of rain and cloud water as well as prognostic cloud condensation nuclei for warm-rain processes, which is relevant for warm and moist precipitation environments (Lim and Hong, 2010). The selected Shin-Hong PBL scheme includes scale dependency for vertical transport in the convective boundary layer while retaining YSU-type treatment for stable boundary-layer and free-atmospheric mixing (Shin and Hong, 2015). This feature is consistent with the need to represent vertical exchange of heat, moisture, and momentum in coastal regions where sea-land thermal contrast and moist boundary-layer processes affect near-surface wind and humidity. The Grell 3 cumulus scheme is an improved Grell-Devenyi type mass-flux approach and can represent sub-grid convective effects through ensemble-based convective assumptions (Grell, 1993; Grell and Devenyi, 2002). Its selection in the Southeast domain is therefore consistent with the higher weight assigned to precipitation in the cumulus evaluation and with previous studies showing that convection and microphysics parameterizations strongly influence precipitation and cloud properties over Asia and East Asian summer monsoon rainfall (Liu et al., 2018; Yang et al., 2015). The common selection of GFDL radiation and Noah LSM in both regions further suggests that a balanced treatment of radiative heating and land-surface fluxes is important for the multivariable evaluation, although their specific roles differ between the arid inland and humid coastal climate backgrounds.

Section 4.2 Physical interpretation and practical meteorological relevance (page 32, lines 707-708):

These physical interpretations should be treated as mechanism-informed explanations of the statistical selection results, not as direct process-attribution evidence.

Comment 1.5

The discussion suggests that the dynamic weighting method is transferable to multi-model evaluations. However, weighting distinct models with structural differences presents different challenges compared to selecting parameterization schemes within the unified WRF framework. The authors should expand this section to clarify how the methodology would be adapted for multi-model ensembles, specifically addressing how it accounts for structural uncertainties versus parameter perturbations.

Response 1.5:

We thank the reviewer for highlighting this important conceptual distinction. We agree that extending the performance-based weighting framework from WRF parameterization-scheme

evaluation to multi-model ensembles requires an explicit adaptation of the interpretation of the weighting components. In the present study, all ensemble members are generated within a unified WRF framework, sharing the same dynamical core, model configuration, domain setting, boundary conditions, and evaluation variables. Under this controlled setting, 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 a multi-model ensemble, however, the same statistic would reflect differences among models that may arise from model structures, numerical schemes, spatial resolutions, physical process representations, tuning strategies, and model genealogy. We therefore agree that the distinction between parameter perturbation uncertainty and structural uncertainty needs to be stated explicitly.

Following the reviewer's suggestion, we added a new discussion subsection (Sect. 4.4) to explain how the method could be adapted as a variable-level evaluation component in future multi-model applications. We now clarify that, in a multi-model context, the coefficient of variation of absolute errors should be interpreted as variable-level inter-model spread rather than sensitivity to a specific physical process. With this adapted interpretation, the two weighting components retain a useful diagnostic role: the mean relative error identifies variables with large systematic simulation errors, whereas the coefficient of variation identifies variables for which different models disagree strongly. Combining these two components can help identify variables that are both difficult to simulate and highly uncertain across models. This is particularly relevant for multivariable evaluation because variables such as precipitation or near-surface wind may be masked by more consistently simulated variables when all variables are assigned equal weights.

We further expanded the discussion to define the value and boundary of this possible extension. The proposed method can help preserve the visibility of high-error and high-spread variables in a multivariable evaluation, but it is not intended to replace model-level treatment of structural uncertainty. We now state that future applications to multi-model ensembles should embed this variable-level weighting module within a broader hierarchical evaluation framework, including harmonization of model outputs and, where appropriate, model-level performance weighting, model-independence weighting, model-family grouping, or uncertainty decomposition. We have also stated that testing this variable-level weighting module in multi-model ensembles is an important direction for future work.

Corresponding revisions:

[Section 4.4 Potential extension of the performance-based weighting framework to variable-level evaluation in multi-model \(page 34, lines 797-804\):](#)

[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.

Section 4.4 Potential extension of the performance-based weighting framework to variable-level evaluation in multi-model (page 34, lines 809-816):

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 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.

Section 4.4 Potential extension of the performance-based weighting framework to variable-level evaluation in multi-model (page 35, lines 819-826):

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

Comment 1.6

P1, L13: "an objective dynamic weighting method" – I advise caution with the word "objective." The choice of the Softmax function, the choice of the variable set (7 variables), and the choice of the evaluation order are all subjective decisions made by the researchers. "Data-driven" or "Performance-based" might be more accurate terms.

Response 1.6:

We thank the reviewer for this helpful suggestion. We agree that the word "objective" may have made the methodological status of the framework appear too strong; therefore, the revised manuscript now uses the more accurate term "performance-based". Although the weights are calculated from model-observation performance statistics, the choice of evaluated variables, the Softmax transformation, and the prescribed physical-process screening order are explicit methodological choices. We therefore revised the manuscript to avoid implying that the method is free of researcher-defined assumptions.

In response, we changed the title and the core terminology throughout the Abstract, Introduction, Methods, Discussion, Conclusions, and Supplementary Information. The method is now described as a "performance-based multivariable weighting framework" or as the proposed "MRE $\times$ CV weighting", rather than as an "objective dynamic weighting method." We adopted this wording because it more accurately separates two points: the variable weights are derived from model performance against observations, but the evaluation rule is still defined by transparent and reproducible methodological choices.

We also clarified that this terminology change does not remove the methodological contribution of the study. Instead, it frames the contribution more precisely. The revised framework adds a diagnostic weighting layer to the MVIE evaluation: the MRE component identifies variables that remain difficult to simulate under the current WRF configuration, and the absolute-error CV component identifies variables whose errors respond strongly to the candidate schemes in the physical-process group being evaluated. By combining these two components, the method is designed to retain the influence of variables that are both important sources of simulation difficulty and informative for parameterization-scheme discrimination. This is why we use the term "performance-based" rather than "subjective" or "fully objective."

To make these assumptions and the method's distinct role clearer to readers, we added explanatory text in Sect. 2.3.3, added benchmark-method descriptions in Sect. 2.3.4, and added a conceptual comparison in Supplementary Table S1. These revisions show how the proposed MRE $\times$ CV weighting differs from equal weighting, entropy weighting, coefficient-of-variation

weighting, and CRITIC weighting. Finally, because the reviewer specifically noted the ordered evaluation workflow, we also revised the Discussion to state that the prescribed order from microphysics to planetary boundary layer, cumulus, radiation, and land surface processes is a practical, pre-defined screening sequence based on previous WRF sensitivity and physics-ensemble evaluation studies, but is not theoretically unique.

Corresponding revisions:

Title (page 1, lines 1-2):

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

Abstract (page 1, lines 4-5):

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.

Section 2.3.3 Performance-based multivariable weighting framework (page 11, lines 221-222):

The term performance-based is used because both weighting components are derived from the simulated performance of candidate WRF parameterization schemes against observations.

Section 2.3.3 (page 13, lines 272-279):

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.

Section 2.3.4 Benchmark weighting methods (page 14, lines 307-331):

The revised manuscript and Supplementary Information compare equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $\text{MRE} \times \text{CV}$ weighting in terms of weighting basis, information emphasized, and interpretation in WRF parameterization screening.

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 34, lines 777-784):

Nevertheless, the order is not theoretically unique. The proposed weighting framework helps reduce evaluation bias within each step by emphasizing variables with larger errors or stronger scheme-dependent responses, but it does not eliminate the path dependence of the stepwise

[search itself.](#)

Supporting material cited in this response: Table S1

Table S1 Conceptual comparison of the variable weighting methods used in this study

Method	Weighting basis	Main information emphasized	Explicit term for error magnitude	Response term reflecting scheme dependence	Conflict term among variables	Interpretation in WRF parameterization screening
Equal weighting	Fixed identical weights for all variables	No variable differentiation	No	No	No	Baseline assumption that all evaluated variables contribute equally to the integrated score.
Entropy weighting	Entropy of normalized scheme-level error distributions	Discriminating information among candidate schemes	Indirect	No	No	Variables with stronger error dispersion among candidate schemes receive greater influence.
Coefficient-of-variation weighting	Ratio of standard deviation to mean error across candidate schemes	Relative error dispersion	Indirect	No	No	Variables with larger relative variation among candidate schemes receive greater influence.
CRITIC	Contrast intensity multiplied by conflict/non-redundancy among variables	Contrast and non-redundant information	Indirect	No	Yes	Variables receive higher weights when they vary strongly among schemes and are less redundant with other variables.
Proposed $MRE \times CV$ weighting	Product of the Softmax-transformed mean relative error and the Softmax-transformed absolute-error coefficient of variation, followed by normalization	Simulation difficulty and scheme-dependent response	Yes	Yes	No	Variables that are both difficult to simulate and responsive to alternative schemes within the evaluated physical-process group retain greater influence in the multivariable score.

Comment 1.7

P2, L43: "To evaluate the overall performance of the model (Gleckler et al., 2008), proposed..." should be corrected to "To evaluate the overall performance of the model, Gleckler et al. (2008) proposed..."

Response 1.7:

Thank you for the careful observation about the citation syntax. We have corrected the citation of Gleckler et al. (2008) so that the subject and grammatical structure of the sentence are clear.

Corresponding revisions:

[Section 1 Introduction \(page 2, lines 23-25\):](#)

To evaluate the overall performance of the model, Gleckler et al. (2008) proposed the Model Climate Performance Index (MCPI), which quantifies the model's overall bias by calculating the mean relative error across multiple variables.

Comment 1.8

P2, L64: "total of over 14.5 million possible combinations" — While mathematically true, many of these are physically invalid or incompatible. It would be better to cite a realistic number of feasible combinations if possible, or acknowledge this is a theoretical upper bound.

Response 1.8:

Thank you for pointing out that the reported number represents a mathematical maximum rather than the number of feasible WRF configurations. We have revised the Introduction to explicitly state that the value is a theoretical upper bound before physical compatibility constraints are considered, and that it should not be interpreted as the number of physically feasible WRF configurations.

Corresponding revisions:

[Section 1 Introduction \(page 2, lines 35-42\):](#)

The size of the configuration space also creates a practical constraint. In WRF version 4.3, the available schemes yield a theoretical upper bound of more than 14.5 million 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).

Section 1 Introduction (page 2, lines 34-36):

This value should therefore be interpreted as an upper bound rather than the number of physically feasible WRF configurations.

Comment 1.9

P4, L114: "we will conduct" should be changed to "we conducted" or "we conduct" to align with standard academic tense for methodology.

Response 1.9:

Thank you for the reminder about tense consistency in the methodological description. We have changed the relevant future-tense expression to the past tense to reflect that the comparative application had already been conducted.

Corresponding revisions:

Section 1 Introduction (page 2, lines 57-58):

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.

Comment 1.10

P5, L140: "Northwest of China" should be corrected to "Northwest China" for consistency.

Response 1.10:

Thank you for pointing out the inconsistency in the regional name. We have corrected "Northwest of China" to "Northwest China" to keep the terminology consistent.

Corresponding revisions:

Section 2.1.1 Arid region of Northwest China (page 3, lines 78-85):

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).

Comment 1.11

P7, L197: "suffering from severe flooding" should be changed to "experiencing severe flooding" or "affected by severe flooding" to maintain an objective academic tone.

Response 1.11:

Thank you for the suggestion to maintain a more objective academic tone. We have replaced the subjective phrase "suffering from severe flooding" with "was affected by severe flooding" and adjusted the parallel clause so that the sentence is grammatically clearer.

Corresponding revisions:

Section 2.2 Model configuration and datasets (page 5, lines 130-134):

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 was affected by severe flooding.

Comment 1.12

P13, L324: "assigned on the basis of" is wordy and should be condensed to "assigned based on" to improve flow.

Response 1.12:

Thank you for the suggestion to improve the flow of this sentence. We have shortened "assigned on the basis of" to "assigned based on" in the descriptions of both weight components in Fig. 2.

Corresponding revisions:

Section 3.1 Optimal parameterization scheme combination for the arid region of Northwest (page 15, lines 336-343):

Figure 2(a) shows the weights ω_{m1} assigned based on the mean relative error ($\overline{MRE_m}$), where wind speed and precipitation received average weights of 0.45 and 0.24, respectively, while all other variables had weights below 0.01. The higher weights for wind speed and precipitation

indicate that all physical processes in the WRF model have significant errors in simulating these variables, reflecting the model's weaker simulation capability for these variables.

Section 3.1 Optimal parameterization scheme combination for the arid region of Northwest (page 15, lines 340-346):

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 significant differences in the sensitivity of the variables to different physical processes.

Comment 1.13

P24, L565: "simplifies the complex multiplicative combinations... into additive combinations." This phrasing is technically incorrect. The physical processes are coupled nonlinearly (multiplicative nature). The step-wise selection method treats them sequentially (additive nature of the process, not the physics). Please rephrase to avoid implying that the method changes the physical nature of the interactions.

Response 1.13:

Thank you for identifying the potential technical ambiguity in this statement. We have rewritten the relevant discussion to clarify that the stepwise screening procedure evaluates physical-process options sequentially, but it does not convert nonlinear interactions among WRF physical processes into independent additive effects.

Corresponding revisions:

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 33, lines 755-762):

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 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 emphasizes the variables most affected by simulation error and

scheme-dependent response.

Comment 1.14

P25, L598: The sentence "...yields more accurate temperature and wind speed, the dynamic weighting scheme provides more accurate results..."The two clauses are tautological/redundant, making the sentence logically circular.

Response 1.14:

Thank you for pointing out the redundancy and circular wording in this conclusion sentence. We have rewritten the sentence to remove the repeated clause and to state the selected configuration in terms of integrated cMIEI, evaluated agreement with observed meteorological variability, remaining biases, and the distinction between evaluation improvement and further WRF physics development.

Corresponding revisions:

Section 5 Conclusions (page 35, lines 839-841):

Similarly, in 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 wind speed. These remaining biases indicate that the weighting method improves the evaluation and selection of parameterization schemes, whereas further refinement of WRF physical parameterizations remains necessary for representing complex weather systems and extreme events.

Comment 1.15

Figure 2 and Figure 5: The x-axis represents discrete categories (Microphysics, PBL, etc.), yet the values are connected by lines. This implies continuity which does not exist. A bar chart or a scatter plot would be a more appropriate visualization for these weights to avoid misleading the reader.

Response 1.15:

Thank you for noting that connecting values across discrete physical-process categories could

imply a continuity that does not exist. We have redrawn Fig. 2 and Fig. 5 as discrete matrix-style heatmaps, and the revised captions now describe the MRE-based weight, CV-based weight, and final proposed weight without using connecting lines.

Corresponding revisions:

Figure 2 (page 15, lines 329-330):

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.

Figure 5 (page 20, lines 433-434):

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.

Supporting material cited in this response: Figure 2 and Figure 5

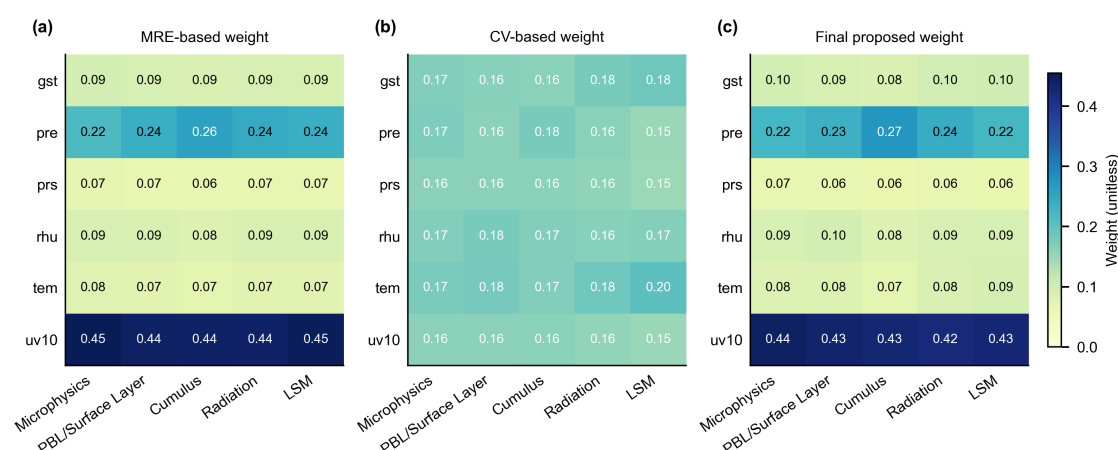


Figure 2. Weight distributions of the variables associated with the different physical processes of the WRF simulation in the arid region of Northwest. (a) MRE-based weight, (b) CV-based weight, and (c) final proposed weight. Weights are unitless.

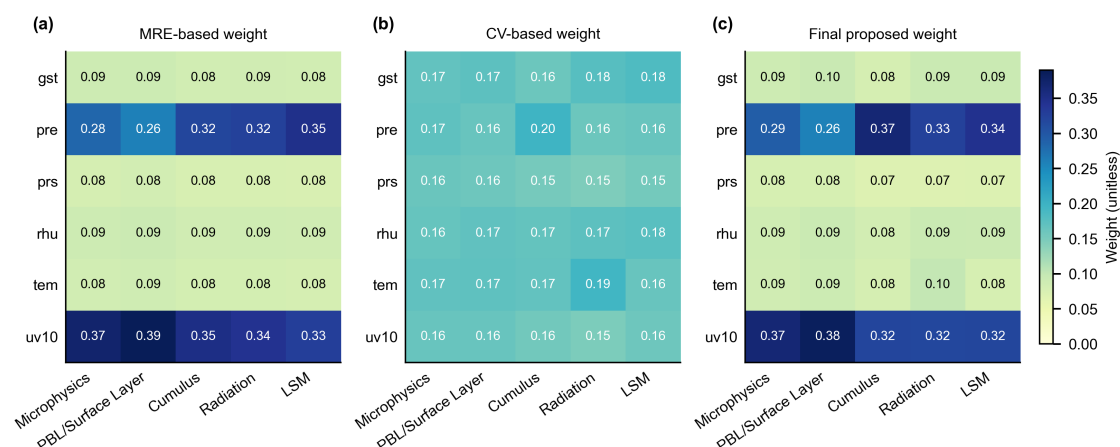


Figure 5. Weight distributions of the variables in the WRF simulations of the different physical processes in the coastal region of Southeast. (a) MRE-based weight, (b) CV-based weight, and (c) final proposed weight. Weights are unitless.

Comment 1.16

Figure 3(a) and Figure 6(a): The color scale for MIEI seems to show "normalized" values. It would be helpful to clarify in the caption what the normalization base is (e.g., normalized by the median of the ensemble?).

Response 1.16:

Thank you for reminding us to specify the normalization basis for the MIEI color scale. We have clarified in the captions of Fig. 3 and Fig. 6 that the normalized MIEI values were normalized by the median MIEI value of the corresponding ensemble.

Corresponding revisions:

Figure 3 (page 17, lines 380-382):

Figure 3. Evaluation results of parameterization schemes in the arid region of Northwest. (a) Normalized MIEI, with MIEI values normalized by the median MIEI value of the corresponding ensemble, and (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 a total of 44 combinations of parametric schemes for all physical processes.

Figure 6 (page 22, lines 1-8):

Figure 6. Evaluation results of parameterization schemes in the coastal region of Southeast. (a) Normalized MIEI, with MIEI values normalized by the median MIEI value of the corresponding ensemble, and (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 a total of 44 combinations of parametric schemes for all physical processes.

ing ensemble, and (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 a total of 54 combinations of parametric schemes for all physical processes.

References cited in the responses to Reviewer 1

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.

Dai, D., Chen, L., Ma, Z., and Xu, Z.: Evaluation of the WRF physics ensemble using a multivariable integrated evaluation approach over the Haihe river basin in northern China, *Clim. Dyn.*, 57, 557-575, doi:10.1007/s00382-021-05723-x, 2021.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

Grell, G. A.: Prognostic evaluation of assumptions used by cumulus parameterizations, *Mon. Weather Rev.*, 121, 764-787, 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.

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.

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.

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.

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.

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.

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.

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.

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.

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. 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.

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.

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.

Response to Reviewer 2

General Comments

The manuscript addresses a relevant problem in WRF parameterization scheme selection, namely how to evaluate and compare different parameterization combinations using multiple meteorological variables. However, the main methodological contribution is not yet convincing. The proposed “dynamic” weighting method appears to be a post hoc error-based weighting scheme, but the manuscript does not sufficiently demonstrate its novelty, theoretical basis, robustness, or advantage over existing objective weighting methods. Major methodological clarification and additional validation are therefore needed. My detailed comments are as follows:

Response to General Comments:

We sincerely thank the reviewer for this thoughtful and constructive overall assessment of our manuscript. We especially appreciate that the reviewer identified the central methodological issues that needed to be strengthened before the contribution of the study could be evaluated convincingly. Although this comment is presented as a general assessment rather than a single specific point, we treated it as a guiding comment for the entire revision. In response, we carefully re-examined the framing, terminology, mathematical definition, benchmark comparisons, robustness tests, and interpretation of the proposed weighting framework.

We agree that the original manuscript needed to define the contribution and scope of the method more clearly, particularly the use of the terms “objective” and “dynamic”, the comparison with other weighting methods, the Softmax normalization, the multiplicative combination of the two weight components, and the stability assessment of the derived weights. We have therefore revised the manuscript substantially rather than treating this as a wording issue.

The revised manuscript now frames the method as a performance-based multivariable weighting framework, not as a fully objective or simulation-online dynamic method. This change is important conceptually: the weights are calculated from model-observation performance statistics, but the choice of variables, the MRE and CV diagnostic components, the Softmax transformation, the multiplicative combination, and the stepwise screening order are all explicit methodological choices. We now state this boundary in the Methods and Discussion, and we removed claims implying that the method provides a uniquely objective weighting framework.

We also strengthened both the theoretical and empirical basis of the method. The MRE component is used to represent variable-level simulation difficulty, whereas the absolute-error CV component is used to represent the scheme-dependent response of each variable within the evaluated physical-process group. The multiplicative rule is now explained as a non-compensatory joint constraint: a variable receives a high final weight only when it is both difficult to simulate and informative for distinguishing candidate parameterization schemes. To test this choice, we

added an equally weighted additive-combination sensitivity analysis and reported the results in Supplementary Table S10.

Finally, we added multiple empirical checks requested by the reviewer. We compared the proposed $\text{MRE} \times \text{CV}$ method with equal weighting, entropy weighting, coefficient-of-variation weighting, and CRITIC weighting; added conceptual comparison in Supplementary Table S1; added complete process-level weight comparisons in Figs. S3-S4 and the main-text Fig. 8; expanded the screening-year performance comparison to five methods in Table 3; expanded the independent 2010 validation to five methods in Table 4; added normalization sensitivity analysis in Fig. S5; added sampling and variable-set stability analyses in Fig. S6 and Tables S8-S9; added component-combination sensitivity analysis in Table S10; added inter-variable dependence and no-GST redundancy checks in Fig. S7 and Table S11; added a pathway analysis linking variable weights, rank reversals, and final scheme selection in Table S12; and added scheme-ranking stability analysis in Fig. S8 and Table S13. These revisions show that the proposed framework is not merely a post hoc error-weighting replacement for equal weighting. Its contribution is to make WRF parameterization-scheme evaluation more transparent and diagnostically focused by jointly incorporating variable-level simulation difficulty and scheme-dependent response under explicit, reproducible methodological choices.

The above summarizes our overall response to the reviewer's general methodological concern. Our detailed point-by-point responses to the reviewer's specific comments are provided below.

In the point-by-point responses below, our responses are shown in regular black font, [whereas the corresponding revisions made in the manuscript are shown in blue regular font](#).

Comment 2.1

1. The comparison with equal weighting alone is too weak to establish methodological superiority. Weighting variables according to error magnitude or error dispersion is a common idea in multivariable evaluation, and the manuscript should demonstrate how the proposed scheme differs conceptually and quantitatively from established objective weighting approaches (e.g., entropy weighting, coefficient-of-variation weighting, CRITIC-type methods, and other sensitivity-based evaluation frameworks).

Response 2.1:

We thank the reviewer for raising this critical methodological concern. We agree that a comparison with equal weighting alone did not sufficiently show the conceptual and quantitative differences between the proposed method and existing statistical weighting approaches. This comment is also consistent with the reviewer's broader concern that the revised manuscript should more clearly explain the theoretical basis, novelty, robustness, and advantages of the

proposed weighting scheme relative to existing statistical weighting approaches. In response, we substantially revised the manuscript and added new benchmark weighting methods, weight-structure comparisons, parameterization-scheme selection comparisons, and independent-year validation analyses.

First, we clarified that the proposed method is not intended to be merely a post hoc error-based replacement for equal weighting. The method is designed to address a specific scientific problem in WRF parameterization-scheme evaluation: variables with large simulation errors can be diluted in a multivariable average, and different variables respond unevenly to different physical parameterization processes. The proposed weighting therefore combines two physically interpretable evaluation components. The mean relative error (MRE) represents variable-level simulation difficulty, quantified by relative simulation error, whereas the coefficient of variation of absolute errors represents the scheme-dependent response of each variable within the physical-process group being evaluated. After multiplication and normalization, the final weight gives greater influence to variables that are both difficult to simulate and useful for distinguishing the parameterization choices under evaluation. By contrast, entropy weighting and coefficient-of-variation weighting mainly emphasize the dispersion of variable errors among candidate schemes, and CRITIC weighting emphasizes contrast intensity and non-redundancy among variables. These methods are useful general benchmark weighting approaches, but they do not explicitly combine simulation difficulty and scheme-dependent response, which are the two elements targeted by the proposed WRF parameterization-screening framework. We added a new Sect. 2.3.4 to describe entropy weighting, coefficient-of-variation weighting, and CRITIC weighting, and we added Table S1 at the beginning of the Supplementary Information to compare the conceptual basis of the five weighting methods.

Second, we added a quantitative comparison of the weights obtained from the five methods and included it as Fig. 8 in the main text. The results show that the methods produced clearly different weight structures. Equal weighting assigns identical weights to all six evaluated variable groups. CRITIC weighting produced weights close to a uniform distribution in the present dataset, with average weights ranging from 0.145 to 0.185 in Northwest China and from 0.142 to 0.199 in Southeast China. Because CRITIC explicitly includes an inter-variable conflict/non-redundancy term, it provides a redundancy-sensitive benchmark. We further examined the two CRITIC components and found that the contrast-intensity term showed only moderate differences among variables, whereas the conflict/non-redundancy term introduced some variable differences but did not create a highly concentrated weight structure. Therefore, the final CRITIC weights remained close to equal weighting in this dataset. Entropy weighting and coefficient-of-variation weighting produced more concentrated and mutually similar weight patterns because both primarily respond to the dispersion of variable errors across candidate schemes. For example, in Northwest China, both methods assigned high average weights to TEM, reaching 0.448 for entropy weighting and 0.356 for coefficient-of-variation weight-

ing, while assigning low weights to PRS and UV10. In contrast, the proposed $\text{MRE} \times \text{CV}$ method consistently assigned high weights to PRE and UV10, with average weights of 0.240 and 0.431 in Northwest China and 0.317 and 0.342 in Southeast China, respectively. This difference demonstrates that the proposed weights are not controlled only by error dispersion, but by the joint effect of simulation difficulty and scheme-dependent response. The complete process-level weight heatmaps for the five methods were also added to the Supplementary Information as Figs. S3-S4.

Third, we compared how these different weighting methods affected the final parameterization-scheme selection, not only the weights themselves. We added Table 3 to compare the configurations selected by equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $\text{MRE} \times \text{CV}$ weighting in the two screening years. In Northwest China in 2007, the proposed method achieved the lowest cMIEI (1.166), compared with 1.186 for equal weighting and CRITIC weighting, 1.198 for entropy weighting, and 1.207 for coefficient-of-variation weighting. In Southeast China in 2009, the proposed method also achieved the lowest cMIEI (1.008), slightly lower than coefficient-of-variation weighting (1.009), equal weighting and CRITIC weighting (1.017), and entropy weighting (1.020). These results show that established benchmark weighting approaches can lead to different WRF parameterization-screening outcomes, and that the proposed method produced the most favorable integrated ranking in the present cMIEI-based screening experiment.

Fourth, to address the concern about robustness and methodological advantage, we also expanded the 2010 independent extreme-year validation from a two-method comparison to a five-method comparison. We removed the original 2010 Taylor diagrams from the main text and replaced them with Table 4. In the 2010 validation for Northwest China, the proposed $\text{MRE} \times \text{CV}$ method retained the lowest cMIEI (1.310), compared with 1.313 for equal weighting and CRITIC weighting, 1.324 for entropy weighting, and 1.322 for coefficient-of-variation weighting. In Southeast China in 2010, the proposed method again achieved the lowest cMIEI (1.008), compared with 1.017 for equal weighting and CRITIC weighting, 1.017 for entropy weighting, and 1.034 for coefficient-of-variation weighting. Thus, the added analyses show not only that the proposed method differs from established benchmark weighting methods in the screening years, but also that the selected configurations retained favorable integrated performance in an independent extreme climate year.

It is worth noting that practical WRF applications can differ substantially in study region, evaluated variables, and decision objectives. Therefore, the choice of a weighting method should be guided by the user's evaluation objective, application context, and the conceptual definition of variable importance embedded in each weighting method. The main value of the benchmark comparison is to show that, under the same WRF evaluation workflow, different weighting concepts can lead to different parameterization-scheme screening outcomes. These outcomes

should be interpreted as framework-dependent selections with different evaluation meanings. Selecting a weighting method according to the specific application need is therefore an important consideration when applying multivariable WRF evaluation frameworks.

Finally, we revised the Abstract, Introduction, Discussion, and Conclusion to align the claims with this expanded evidence. We no longer claim that the proposed method is superior in every individual statistical metric or that it directly improves the WRF physical parameterizations. Instead, we now define it as a performance-based multivariable weighting framework for WRF parameterization-scheme evaluation. Its contribution is that it jointly incorporates variable-level simulation difficulty and scheme-dependent response into the weighting calculation, and its conceptual and quantitative differences from equal weighting, entropy weighting, coefficient-of-variation weighting, and CRITIC weighting are now explicitly demonstrated.

Corresponding revisions:

Abstract (page 1, lines 7-9):

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 wind speed, 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.

Section 2.3.4 Benchmark weighting methods (page 14, lines 307-331):

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. These methods were recalculated within the same stepwise WRF parameterization-screening workflow and compared with both equal weighting and the proposed $\text{MRE} \times \text{CV}$ weighting. The proposed $\text{MRE} \times \text{CV}$ 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.

Section 3.5 Comparison with established statistical weighting methods (page 24, lines 531-538):

Figure 8 shows that the five weighting methods produced different average variable-weight structures. 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 0.142 to 0.199 in 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, especially in Southeast China, where both methods assigned larger weights to PRE and UV10. The proposed $MRE \times CV$ method produced a different weight structure. It consistently assigned larger weights to PRE and UV10, but this pattern came from the joint effect of the MRE and absolute-error CV components rather than from dispersion alone.

Table 3 (page 26, lines 570-574):

Performance comparison of the configurations selected by five weighting methods in the screening years. In Northwest China in 2007, the proposed method achieved $cMIEI = 1.166$, compared with 1.186 for equal weighting and CRITIC weighting, 1.198 for entropy weighting, and 1.207 for coefficient-of-variation weighting. In Southeast China in 2009, the proposed method achieved $cMIEI = 1.008$, compared with 1.009 for coefficient-of-variation weighting, 1.017 for equal weighting and CRITIC weighting, and 1.020 for entropy weighting.

Section 3.6 Independent 2010 validation of the five weighting methods (page 27, lines 585-590):

The 2010 validation results are summarized in Table 4. In Northwest China, the proposed $MRE \times CV$ method retained the lowest $cMIEI$ (1.310), slightly lower than equal weighting and CRITIC (1.313), entropy weighting (1.324), and coefficient-of-variation weighting (1.322). In Southeast China, the proposed method again produced the lowest $cMIEI$ (1.008), compared with 1.017 for equal weighting and CRITIC, 1.017 for entropy weighting, and 1.034 for coefficient-of-variation weighting. This independent extreme-year test provides additional evidence that the proposed weighting strategy can retain its integrated performance advantage outside the screening year.

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 29, lines 607-614):

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 have a physical evaluation meaning, because they identify variables that are both difficult to simulate and influential for distinguishing the parameterization choices under evaluation.

Supporting Information, Table S1:

Conceptual comparison of the variable weighting methods used in this study. The table compares equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $\text{MRE} \times \text{CV}$ weighting in terms of weighting basis, information emphasized, explicit error-magnitude term, process-sensitivity term, and inter-variable conflict term.

Supporting material cited in this response: Table S1, Figure 8, Figs. S3–S4, Table 3, and Table 4

Table S1 Conceptual comparison of the variable weighting methods used in this study

Method	Weighting basis	Main information emphasized	Explicit term for error magnitude	Response term reflecting scheme dependence	Conflict term among variables	Interpretation in WRF parameterization screening
Equal weighting	Fixed identical weights for all variables	No variable differentiation	No	No	No	Baseline assumption that all evaluated variables contribute equally to the integrated score.
Entropy weighting	Entropy of normalized scheme-level error distributions	Discriminating information among candidate schemes	Indirect	No	No	Variables with stronger error dispersion among candidate schemes receive greater influence.
Coefficient-of-variation weighting	Ratio of standard deviation to mean error across candidate schemes	Relative error dispersion	Indirect	No	No	Variables with larger relative variation among candidate schemes receive greater influence.
CRITIC	Contrast intensity multiplied by conflict/non-redundancy among variables	Contrast and non-redundant information	Indirect	No	Yes	Variables receive higher weights when they vary strongly among schemes and are less redundant with other variables.
Proposed $MRE \times CV$ weighting	Product of the Softmax-transformed mean relative error and the Softmax-transformed absolute-error coefficient of variation, followed by normalization	Simulation difficulty and scheme-dependent response	Yes	Yes	No	Variables that are both difficult to simulate and responsive to alternative schemes within the evaluated physical-process group retain greater influence in the multivariable score.

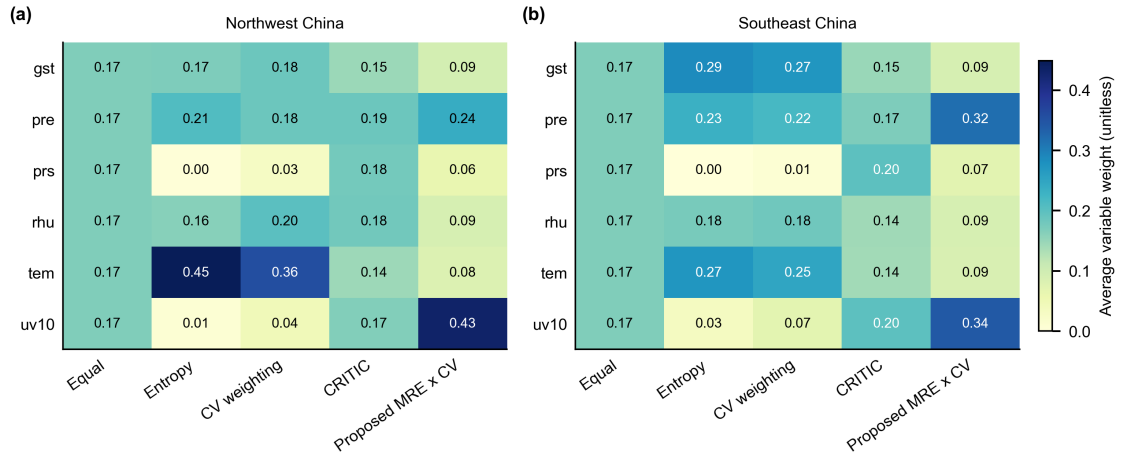


Figure 8. Comparison of average variable weights derived from equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed MRE $\times$ CV weighting. Values represent averages across the five physical-process evaluation stages. (a) Arid region of Northwest China; (b) coastal region of Southeast China. Weights are unitless.

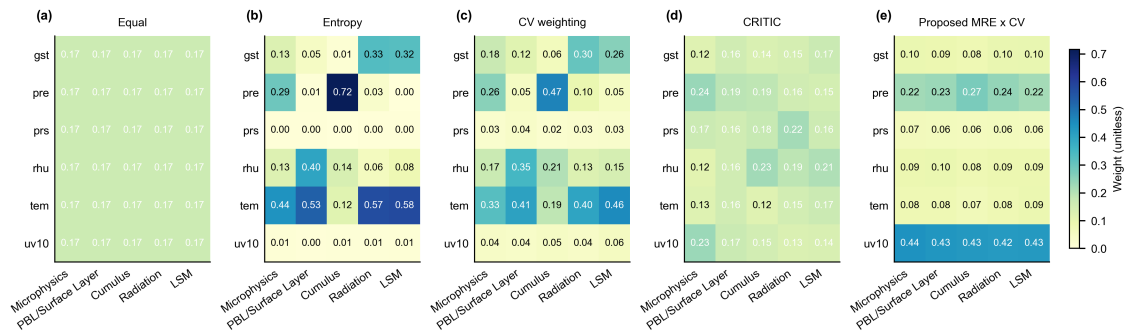


Figure S3. Complete process-level variable weights derived from equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed MRE $\times$ CV weighting in the arid region of Northwest China. Weights are unitless and were calculated for each physical-process evaluation stage.

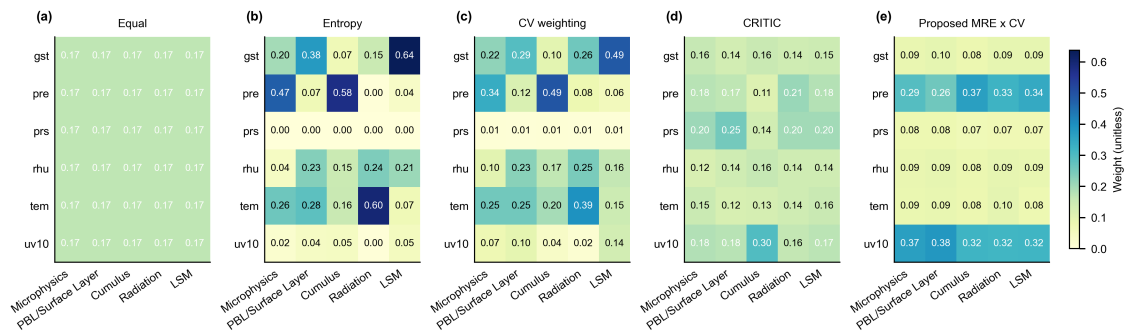


Figure S4. Complete process-level variable weights derived from equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $\text{MRE} \times \text{CV}$ weighting in the coastal region of Southeast China. Weights are unitless and were calculated for each physical-process evaluation stage.

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.

Region	Screening year	Weighting method	cMIEI	cVSC	cRMSL	Rank by cMIEI
Northwest China	2007	Equal weighting	1.186	0.321	1.167	2
		Proposed $\text{MRE} \times \text{CV}$	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	1.017	0.503	1.186	3
		Proposed $\text{MRE} \times \text{CV}$	1.008	0.505	1.086	1
		Entropy weighting	1.020	0.495	1.081	5
		CV weighting	1.009	0.505	1.008	2
		CRITIC weighting	1.017	0.503	1.186	3

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.

Region	Validation year	Weighting method	cMIEI	cVSC	cRMSL	Rank by cMIEI
Northwest China	2010	Equal weighting	1.313	0.190	1.239	2
		Proposed $\text{MRE} \times \text{CV}$	1.310	0.181	1.137	1
		Entropy weighting	1.324	0.225	1.404	5
		CV weighting	1.322	0.219	1.392	4
		CRITIC weighting	1.313	0.190	1.239	2
Southeast China	2010	Equal weighting	1.017	0.504	1.065	3
		Proposed $\text{MRE} \times \text{CV}$	1.008	0.508	1.067	1
		Entropy weighting	1.017	0.500	1.063	2
		CV weighting	1.034	0.486	0.988	5
		CRITIC weighting	1.017	0.504	1.065	3

Comment 2.2

2. The meaning of “dynamic” is unclear. The weights seem to be calculated after the simulations are completed, rather than being dynamically updated during model simulation or evaluation.

Response 2.2:

We thank the reviewer for pointing out this ambiguity. We agree that the original wording did not sufficiently distinguish the evaluation-stage weighting procedure from online model adaptation. We clarify that the proposed weighting framework is not simulation-online: the weights are not updated during WRF numerical integration, do not feed back to the model state, physical parameterizations, or time-integration procedure, and are not adjusted in real time at each time step, day, or station.

At the same time, the framework does not use one fixed set of weights for all simulations or all evaluation stages. It is better described as stage-adaptive post-simulation evaluation weighting. For each physical-process screening stage, the candidate schemes in that process group are first simulated under the current background configuration. After these candidate simulations have been completed, the MRE and CV statistics are calculated from the error structure of that stage-specific candidate set, and the variable weights are recalculated before ranking the schemes in that stage. The selected scheme is then carried forward as part of the background

configuration for the next physical-process screening stage. Therefore, the variation in weights arises from the evaluation workflow itself: the weight structure can change with the evaluated variables, physical-process group, study region, evaluation period, and candidate-scheme error structure. Although the weights do not feed back into WRF integration, they affect the stage-wise parameterization-scheme selection and can influence the subsequent stepwise screening path.

Because the term "dynamic" could obscure this distinction, we revised the manuscript terminology. We removed "dynamic" from the title and core method name and now describe the method as a performance-based multivariable weighting framework. We also revised the manuscript and Supplementary Information to use "the proposed $\text{MRE} \times \text{CV}$ weighting" where the specific weighting method is meant. In Sect. 2.3.3, we explicitly clarify that the weights are not updated online during WRF model integration, but are recalculated after the candidate simulations in each physical-process group have been completed and before the schemes in that group are ranked.

Corresponding revisions:

Title (page 1, lines 1-2):

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

Abstract (page 1, lines 4-5):

To address this issue, this study proposes a performance-based multivariable weighting framework for WRF parameterization-scheme evaluation.

Section 2.2 Model configuration and datasets (page 5, lines 116-118):

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 $\text{MRE} \times \text{CV}$ weighting and equal weighting methods.

Section 2.3.3 Performance-based multivariable weighting framework (page 11, lines 222-223):

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.

Supporting Information, Tables S4 and S6:

Table S4. Summary of simulation experiment on the proposed $\text{MRE} \times \text{CV}$ weighted combination of parameterization schemes in the arid region of Northwest China. Table S6. Summary of

simulation experiment on the proposed $\text{MRE} \times \text{CV}$ weighted combination of parameterization schemes in the coastal region of Southeast China.

Comment 2.3

3. The objectivity of the method is overstated. Although the weights are calculated from data, several key components of the method are still based on subjective methodological choices, including the selection of MRE and CV as the two weighting metrics, the use of the Softmax transformation, and the multiplicative combination of the two weight components. These choices require stronger theoretical or empirical justification before the method can be presented as a fully objective weighting framework.

Response 2.3:

We thank the reviewer for this important methodological comment. We agree that the term “objective” in the original manuscript needed to be further qualified to avoid suggesting that the method is a fully objective weighting framework without any researcher-defined methodological choices. The weights are calculated from model-observation performance data, but several elements of the framework, including the choice of MRE and CV, the Softmax transformation, the multiplicative combination, the evaluated variable set, and the ordered screening procedure, are explicit methodological choices. We therefore revised the title and manuscript terminology to avoid presenting the method as a fully objective weighting framework.

In the revised manuscript, we define the method as a performance-based multivariable weighting framework. We do not present MRE or CV themselves as new statistics. Instead, the methodological contribution is to assign these two familiar error diagnostics separate and complementary roles in WRF parameterization-scheme screening. The first role is to identify variables that remain difficult to simulate in a given region and modelling configuration. For this purpose, MRE is used because it expresses model error relative to the observed magnitude of each variable, making variables with different units and scales more comparable in a multivariable evaluation. In this framework, a large MRE indicates that a variable is a limiting variable for the current WRF evaluation, rather than simply another component in an arithmetic average.

The second role is to identify whether a variable is informative for distinguishing the candidate schemes within the physical-process group being evaluated. For this purpose, the coefficient of variation is calculated from the absolute errors of the candidate schemes for the same variable and process group. Mathematically, this CV is a normalized measure of error dispersion among schemes. In the WRF screening context, a low CV means that candidate schemes produce similar errors for that variable, so the variable provides limited information for choosing among schemes even if its mean error is large. A high CV means that the variable error changes sub-

stantially across candidate schemes and therefore carries scheme-discrimination information. This is why we interpret the absolute-error CV as the scheme-dependent response of a variable, not merely as a generic dispersion statistic.

The combination of MRE and CV therefore reflects a specific scientific choice in the present evaluation: the multivariable score should emphasize variables that are both difficult to simulate and useful for distinguishing the parameterization choices under evaluation. A variable with high MRE but low CV may represent a general model bias that is not strongly affected by the scheme options in the current stage; a variable with low MRE but high CV may distinguish schemes but may not be a dominant source of overall simulation difficulty. The proposed $\text{MRE} \times \text{CV}$ weighting is designed to retain the influence of variables that satisfy both conditions. This is the main methodological distinction from entropy weighting and coefficient-of-variation weighting, which mainly emphasize dispersion among candidate schemes, and from CRITIC weighting, which additionally considers inter-variable conflict but does not explicitly represent variable-level simulation difficulty.

We also strengthened the explanation of the Softmax transformation and the multiplicative combination. Softmax is now described as a transparent methodological choice that maps the two diagnostic components to positive relative weights while preserving their rank order and avoiding a hard threshold for retaining or excluding variables. We also made the final normalization explicit, so that the final variable weights sum to one. For the multiplicative rule, we added a non-compensatory interpretation: a variable receives a high final weight only when both diagnostic components support it. This differs from an additive rule, where a high value in one component can partly compensate for a low value in the other. To test whether this algebraic choice controlled the conclusions, we added an equally weighted additive-combination sensitivity analysis.

To support this revised rationale empirically, we added several checks rather than relying only on conceptual explanation. First, we compared the proposed $\text{MRE} \times \text{CV}$ method with equal weighting, entropy weighting, coefficient-of-variation weighting, and CRITIC weighting, and added a conceptual comparison in Supplementary Table S1. Second, we added weight-structure comparisons in the main-text Fig. 8 and the complete process-level matrices in Figs. S3-S4. Third, we added sensitivity tests for normalization choice in Fig. S5, sampling and variable-set stability in Fig. S6 and Tables S8-S9, and component-combination choice in Table S10. Fourth, we expanded the screening-year and independent 2010 validation comparisons to five weighting methods in Tables 3 and 4. These analyses do not imply that the proposed rule is universally optimal, but they support the more limited claim that the method has a distinct diagnostic meaning and produces traceable screening outcomes in the present WRF parameterization-screening experiment.

We therefore revised the claim in a way that separates methodological contribution from overstatement. The revised manuscript presents the method as an independent and reproducible di-

agnostic weighting rule for multivariable model evaluation: it converts two evaluation-relevant properties of the simulations, variable-level difficulty and within-process scheme-dependent response, into stage-specific variable weights, and tests this rule against four benchmark weighting strategies and several sensitivity analyses. In this study, cMIEI is used as the integrated evaluation index, but the weighting rule is not conceptually tied to cMIEI or to MVIE and can be coupled with other multivariable evaluation indices that retain variable-level performance information. Thus, the contribution is not a claim of complete objectivity, but a methodological extension that makes WRF parameterization screening more diagnostic, comparable, and reproducible than an equal-weighted or single-dispersion weighting setup.

Corresponding revisions:

Title (page 1, lines 1-2):

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

Section 2.3.3 Performance-based multivariable weighting framework (page 11, lines 221-222):

The term performance-based is used because both weighting components are derived from the simulated performance of candidate WRF parameterization schemes against observations.

Section 2.3.3 Performance-based multivariable weighting framework (page 11, lines 228-231):

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. 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.

Section 2.3.3 (page 13, lines 272-279):

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.

Section 2.3.4 Benchmark weighting methods (page 14, lines 307-331):

The revised manuscript and Supplementary Information compare equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $MRE \times CV$ weighting in terms of weighting basis, information emphasized, and interpretation in WRF

parameterization screening.

Section 3.5 Comparison with established statistical weighting methods (page 27, lines 571-580):

The additive-combination sensitivity test showed that the additive alternative selected the same stage-level cases as the proposed multiplicative rule in all five Southeast China stages and in four of five Northwest China stages. In the independent 2010 Northwest China validation, the additive-rule configuration had cMIEI = 1.313, close to but slightly higher than the proposed multiplicative-rule configuration (cMIEI = 1.310).

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 29, lines 608-615):

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.

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 29, lines 632-634):

These results support the use of the multiplicative combination for the present evaluation objective, while not implying that a product rule would be universally superior for all model-evaluation contexts.

Supporting Information, Table S1:

The table compares equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, and the proposed $\text{MRE} \times \text{CV}$ weighting in terms of weighting basis, information emphasized, explicit error-magnitude term, scheme-dependence response term, inter-variable conflict term, and interpretation in WRF parameterization screening.

Comment 2.4

4. The manuscript does not consider correlations among the evaluated variables. The authors should discuss how inter-variable dependence may affect the multivariable evaluation and whether any decorrelation or redundancy-control procedure is needed.

Response 2.4:

We thank the reviewer for raising this important methodological issue. We agree that the original manuscript did not explicitly examine correlations among the evaluated variables, and

that this omission could leave uncertainty about whether related variables were counted as independent information in the multivariable evaluation. This concern is especially relevant for near-surface meteorological variables, because physically related variables can share part of the same thermal, moisture, or dynamical signal. We therefore revised the Methods, Results, Discussion, and Supplementary Information to explicitly evaluate inter-variable dependence and to explain why we did not apply a formal decorrelation transform in the main evaluation.

We addressed the issue in four connected steps. First, we added a correlation analysis in Sect. 2.3.3 and Supplementary Fig. S7. We calculated Pearson correlation matrices from two complementary data views. The first used the observed daily station series and describes the physical co-variation among the evaluated variables. The second used the scheme-level standardized-error matrices in the proposed $\text{MRE} \times \text{CV}$ screening workflow and therefore addresses the more direct question of whether the error information used for parameterization-scheme evaluation contains redundant structure. This distinction is important because correlations among observed meteorological variables and correlations among scheme-level evaluation errors do not necessarily have the same interpretation.

Second, we reported the resulting dependence structure in Sect. 3.5 and Supplementary Fig. S7. 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. This result confirms that the two temperature-related variables are strongly related and that the concern about possible repeated influence is legitimate. At the same time, the result does not by itself mean that one variable is redundant. GST and TEM represent different diagnostic aspects of the near-surface thermal system. GST is more directly connected to surface energy balance, land-surface exchange, and radiative heating or cooling at the ground surface, whereas TEM reflects the near-surface air-temperature response after surface forcing is coupled with surface-layer and boundary-layer mixing, radiative forcing, and transport. Thus, two correlated variables may still have different error responses to different physical-process options. This is exactly the type of information that motivates the absolute-error CV component: it evaluates whether each variable's error varies among the candidate schemes within the physical-process group being evaluated, instead of assuming that physically related variables carry identical screening information.

Third, we used CRITIC as a redundancy-sensitive benchmark. CRITIC explicitly includes an inter-variable conflict/non-redundancy term (Supplementary Table S1), so it provides a useful comparison for evaluating whether adding a non-redundancy component would strongly alter the weight structure. In the present dataset, the CRITIC average weights remained close to a uniform distribution in Fig. 8, indicating that this benchmark did not overturn the main weight-structure evidence.

Fourth, we performed a targeted no-GST sensitivity test, while retaining TEM as the standard

near-surface air-temperature variable. This test directly addresses whether the strongest correlated pair, GST and TEM, controls the final stage-level selections. The selected stage-level case was unchanged in 8 of the 10 region-stage comparisons. The two changes occurred in Southeast China, where the PBL/surface-layer choice changed from Shin-Hong + MM5 to UW + MYJ and the land-surface choice changed from Noah LSM to NoahMP. This pattern is physically interpretable rather than anomalous. GST represents the surface thermal state and is directly linked to surface energy balance, land-atmosphere fluxes, and radiative heating or cooling at the ground surface. TEM, although strongly correlated with GST, represents 2 m near-surface air temperature after surface forcing has been modulated by surface-layer exchange, PBL turbulent mixing, radiation, and transport. Therefore, removing GST can affect not only the land-surface stage, but also the PBL/surface-layer stage. In the humid coastal Southeast domain, removing the direct constraint on surface thermal state makes the integrated evaluation depend more on the combined responses of TEM, RHU, UV10, and PRE, which are partly mediated by boundary-layer mixing and near-surface exchange. A change from Shin-Hong + MM5 to UW + MYJ is therefore physically reasonable because PBL/surface-layer schemes control low-level heat, moisture, and momentum exchange. The change from Noah LSM to NoahMP is also reasonable because NoahMP contains a more detailed land-surface process representation, and after GST is removed the land-surface ranking can be driven more by indirect responses in TEM, RHU, PRE, and UV10. Importantly, the Southeast China cumulus and radiation choices remained Grell 3 and GFDL, respectively. Thus, the effect of removing GST was mainly confined to stages most directly linked to near-surface thermal coupling and land-atmosphere exchange, and it did not overturn the core precipitation/convection and radiation-related selection pathway. We also state explicitly that the original and no-GST cMIEI values should not be interpreted as direct cross-set performance differences, because the evaluated variable sets are different. The relevant sensitivity evidence is the overall stability and physically interpretable sensitivity of the selected stage-level cases.

We further clarify that the no-GST experiment should be interpreted as a targeted sensitivity test for the strongest temperature-related dependence, rather than as a complete redundancy-control analysis for all correlated variable pairs. A fuller redundancy assessment would require additional leave-one-variable-out or grouped-variable experiments, which we regard as useful future extensions beyond the scope of the present revision.

These additions also clarify why we did not apply formal decorrelation or remove a variable in the main evaluation. In this study, the evaluated variables are not generic predictors in a regression model, but physically defined WRF output variables used to diagnose parameterization-scheme performance. Applying a decorrelation or orthogonalization transform would change the evaluation target from interpretable meteorological quantities to statistical components, making the selected schemes less directly interpretable in terms of WRF physical processes. We therefore retained the physically defined variables and evaluated redundancy risk through trans-

parent diagnostics, a CRITIC benchmark, and a targeted variable-removal sensitivity test. This revision addresses variable dependence more explicitly: the revised manuscript no longer implicitly assumes variable independence, reports the main dependence structure, and uses the no-GST sensitivity test to assess whether the principal scheme-selection outcome is affected by the strongest correlated variable pair.

Corresponding revisions:

Section 2.3.3 (page 13, lines 289-291):

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 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 cases. These diagnostics are reported in Fig. S7 and Table S11.

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 29, lines 634-641):

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.

Section 3.5 Comparison with established statistical weighting methods (page 24, lines 534-535):

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.

Section 3.5 Comparison with established statistical weighting methods (page 25, lines 556-563):

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 case 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 S11). 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 cases.

Supporting Information, Fig. S7:

The revised Supplementary Information now reports Pearson correlation matrices calculated from observed daily station series and from the proposed $\text{MRE} \times \text{CV}$ scheme-level standardized-error matrices for both study regions.

Supporting Information, Table S11:

The revised Supplementary Information now reports the no-GST sensitivity test. The selected stage-level case was unchanged in 8 of the 10 region-stage comparisons, with changes in the Southeast China PBL/surface-layer and land-surface stages.

Supporting material cited in this response: Fig. S7 and Table S11

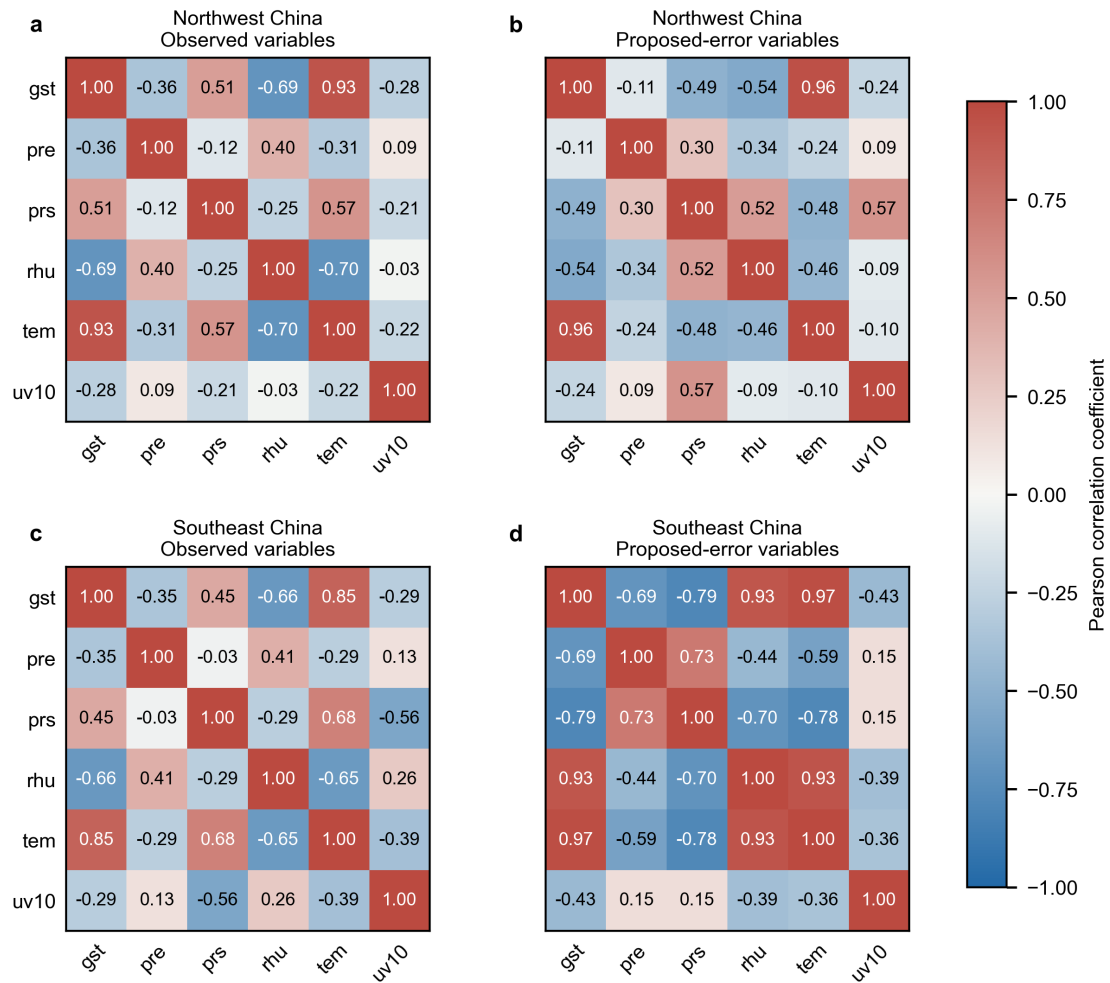


Figure S7. Correlation structure among the evaluated variables. Panels show Pearson correlation matrices calculated from observed daily station series and from the proposed $\text{MRE} \times \text{CV}$ scheme-level standardized-error matrices for Northwest China and Southeast China. Correlation coefficients are unitless. The figure documents both the physical co-variation among observed variables and the redundancy structure in the error information used for multivariable parameterization-scheme evaluation.

Table S11 Sensitivity of selected stage-level schemes to removing GST

Region and year	Physical-process stage	Original best scheme	No-GST best scheme	Stable after removing GST
Northwest China	Cumulus	Tiedtke	Tiedtke	Yes
	Microphysics	P3	P3	Yes
	PBL / surface layer	YSU + MM5	YSU + MM5	Yes
	Radiation	GFDL	GFDL	Yes
	Land surface	Noah LSM	Noah LSM	Yes
Southeast China	Cumulus	Grell 3	Grell 3	Yes
	Microphysics	WDM6	WDM6	Yes
	PBL / surface layer	Shin-Hong + MM5	9 UW + 2 MYJ	No
	Radiation	GFDL	GFDL	Yes
	Land surface	Noah LSM	NoahMP	No

Comment 2.5

5. The stability of the derived weights is not adequately examined. It is unclear whether the variable weights would remain similar if a different year, station subset, time period, or variable set were used. Without such stability tests, the derived weights may reflect sampling noise rather than robust variable importance.

Response 2.5:

We thank the reviewer for raising this important methodological concern. We agree that the original manuscript did not adequately test whether the derived variable weights were stable with respect to station sampling, temporal sampling, the evaluated variable set, or the year background. In response, we added a dedicated weight-stability analysis and revised the Methods, Results, Discussion, and Supplementary Information accordingly. The new material is now reported in Sect. 2.3.3, Sect. 3.5, the Discussion, and Supplementary Information Fig. S6 and Tables S8-S9.

To examine the effects of station subsets and time periods, we performed resampling analyses for each study region and each physical-process evaluation stage. For station stability, we generated 500 bootstrap samples by drawing the original number of stations with replacement while retaining all summer days. For temporal stability, we used 500 moving-block bootstrap samples composed of contiguous 14 d blocks, so that within-summer temporal dependence was partly retained. We also recalculated the weights separately for June, July, and August. For

every resample or monthly subset, the MRE component, the absolute-error CV component, and the final normalized weights were recalculated. Each weight vector was then compared with the full-sample vector using Spearman rank correlation, maximum absolute weight difference, the retention rate of the highest-weight variable, and the retention rate of the unordered set of the two highest-weight variables.

The added results show that the dominant weight structure was stable under the tested station and intra-seasonal temporal perturbations. Across the ten region-stage combinations, the mean Spearman correlation with the full-sample weights ranged from 0.947 to 1.000 for the station bootstrap samples and from 0.944 to 1.000 for the 14 d block bootstrap samples. The lowest lower-tail or monthly-subset correlation was 0.886. PRE and UV10 formed the set of the two highest-weight variable groups in every station bootstrap, time-block bootstrap, and monthly subset, giving a top-two retention rate of 100%. In several Southeast China stages, especially the radiation stage, PRE and UV10 occasionally exchanged the first and second positions. This indicates that the two dominant variables had similar weights in that stage, while the dominant two-variable set remained unchanged.

To examine stability with respect to the evaluated variable set, we further conducted a leave-one-variable-out weight check. After omitting GST, PRE, PRS, RHU, TEM, and UV10 in turn, the retained-variable ordering was preserved in all ten region-stage cases, and the maximum absolute residual between the recomputed weights and the theoretical renormalization was 2.22×10^{-16} . This behavior follows from the normalization structure of the proposed method: because the MRE and absolute-error CV scores are calculated independently for each variable, deleting variable k changes the denominator of the final normalization but does not change the unnormalized scores of the retained variables. Therefore, after deleting variable k , the retained-variable weight satisfies $\omega_m^{(-k)} = \omega_m / (1 - \omega_k)$, and the ratios and ranking among retained variables are preserved. This analysis shows that variable omission changes the absolute normalized shares in a predictable way, while preserving the relative importance among the retained variables.

For the year-related aspect, the revised manuscript presents the available evidence at two levels. First, the 2007 Northwest China and 2009 Southeast China cases both identified PRE and UV10 as the dominant high-weight variables, supporting consistency of the dominant weight structure across the two available region-year settings. Second, the independent 2010 extreme-year validation showed that the configurations selected using the proposed weighting framework retained favorable integrated performance outside the screening years. A fully controlled same-region cross-year weight-stability test would require rerunning the complete stage-wise WRF candidate-scheme ensemble in additional years and recalculating the weights from those simulations; it cannot be obtained by simply post-processing the existing simulations. We therefore added the station, temporal, monthly, and variable-set stability tests that can be directly supported by the current data, and briefly stated the cross-year interpretation boundary in the

Discussion.

Overall, the added analyses examine weight stability from four perspectives: station sampling, intra-seasonal temporal sampling, fixed monthly subsets, and variable-set perturbation. These results indicate that, within the tested station-sampling, temporal-sampling, monthly-subset, and variable-set perturbation ranges, the dominant structure of the $MRE \times CV$ weights is relatively stable and does not show clear evidence of being driven by a single station sample or a particular time-period choice.

Corresponding revisions:

Section 2.3.3 (page 13, lines 276-279):

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.

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 33, lines 745-752):

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.

Section 3.5 Comparison with established statistical weighting methods (page 25, lines 548-556):

The resampling analyses showed that the dominant weight structure was stable to station and intra-seasonal temporal sampling (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 samples. The lowest lower-tail or monthly-subset correlation was 0.886. PRE and UV10 formed the set of

the two highest-weight 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).

Supporting Information, Fig. S6 and Table S8:

The revised Supplementary Information now provides the point-interval plot and exact summary table for station bootstrap, 14 d moving-block bootstrap, and monthly-subset stability, including Spearman rank correlation, maximum absolute weight difference, top-one retention, and top-two retention.

Supporting Information, Variable-set subcomposition stability and Table S9:

The revised Supplementary Information derives and numerically verifies $\omega_m^{(-k)} = \omega_m / (1 - \omega_k)$ after deleting variable k , showing that retained-variable ratios and ordering are preserved and that the residuals are at floating-point precision.

Supporting material cited in this response: Fig. S6 and Tables S8–S9

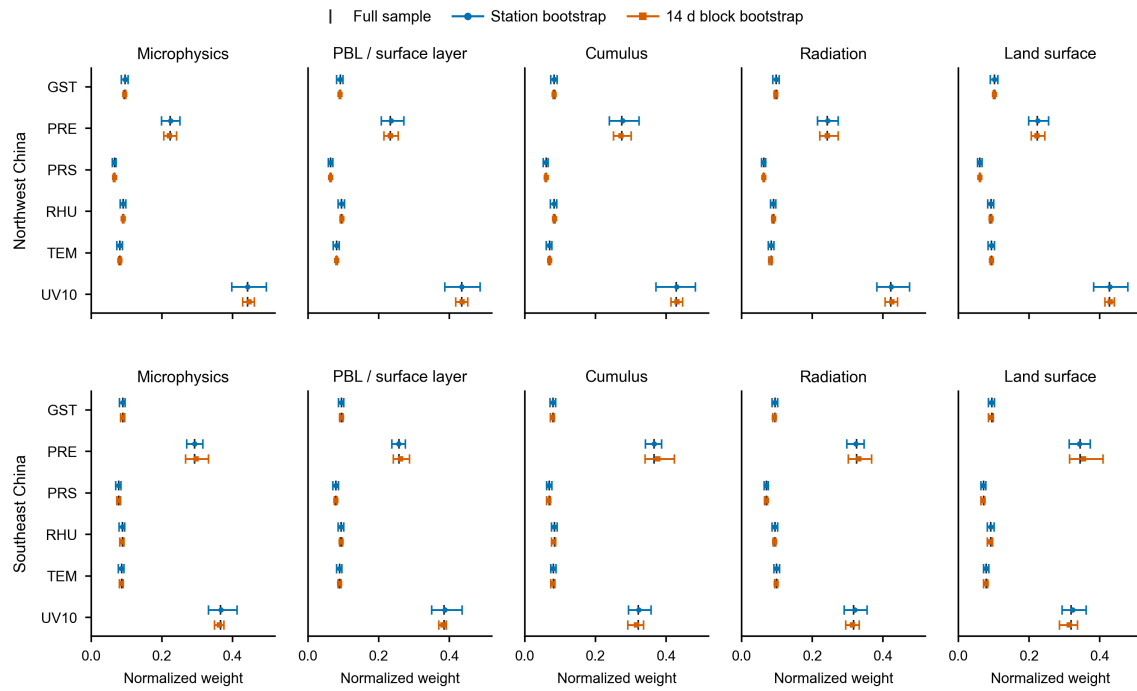


Figure S6. Sampling stability of the normalized variable weights derived using the proposed $MRE \times CV$ method. Rows represent the arid region of Northwest China and the coastal region of Southeast China, and columns represent the five physical-process evaluation stages. Blue circles and orange squares show the mean weights from 500 station-bootstrap and 500 14 d moving-block-bootstrap samples, respectively; horizontal error bars indicate the 2.5th–97.5th percentile intervals. Black vertical markers, repeated on the corresponding station- and time-bootstrap rows, show the full-sample weights. All weights are dimensionless.

Table S8 Exact summary of variable-weight stability under station and temporal sampling

Perturba- tion	Region	Evaluation stage	n	Mean Spearman	min	Mean	max	Top-1 reten- tion	Top-2 reten- tion
Station bootstrap	North- west China	Microphysics	500	1.000	1.000	0.021	0.056	1.000	1.000
		PBL / surface layer	500	0.999	1.000	0.023	0.057	1.000	1.000
		Cumulus	500	0.975	0.943	0.026	0.066	1.000	1.000
		Radiation	500	1.000	1.000	0.021	0.054	1.000	1.000
		Land surface	500	0.984	0.943	0.022	0.056	1.000	1.000
	South- east China	Microphysics	500	0.996	0.943	0.018	0.048	0.998	1.000
		PBL / surface layer	500	0.986	0.943	0.019	0.052	1.000	1.000
		Cumulus	500	0.986	0.943	0.016	0.037	0.942	1.000
		Radiation	500	0.947	0.886	0.017	0.039	0.542	1.000
		Land surface	500	0.979	0.943	0.018	0.043	0.754	1.000
14 d block bootstrap	North- west China	Microphysics	500	1.000	1.000	0.010	0.021	1.000	1.000
		PBL / surface layer	500	1.000	1.000	0.010	0.022	1.000	1.000
		Cumulus	500	0.957	0.943	0.011	0.029	1.000	1.000
		Radiation	500	1.000	1.000	0.012	0.030	1.000	1.000
		Land surface	500	0.992	0.943	0.009	0.022	1.000	1.000
	South- east China	Microphysics	500	0.999	1.000	0.015	0.040	0.994	1.000
		PBL / surface layer	500	0.992	0.943	0.010	0.030	1.000	1.000
		Cumulus	500	0.997	0.943	0.018	0.057	0.988	1.000
		Radiation	500	0.944	0.886	0.014	0.043	0.760	1.000
		Land surface	500	0.987	0.943	0.019	0.065	0.902	1.000
Monthly subsets	North- west China	Microphysics	3	1.000	1.000	0.019	0.033	1.000	1.000
		PBL / surface layer	3	1.000	1.000	0.024	0.039	1.000	1.000
		Cumulus	3	0.981	0.943	0.024	0.039	1.000	1.000
		Radiation	3	1.000	1.000	0.023	0.041	1.000	1.000
		Land surface	3	0.981	0.943	0.018	0.037	1.000	1.000
	South- east China	Microphysics	3	1.000	1.000	0.017	0.023	1.000	1.000
		PBL / surface layer	3	0.981	0.943	0.006	0.009	1.000	1.000
		Cumulus	3	0.962	0.943	0.019	0.028	1.000	1.000
		Radiation	3	0.943	0.886	0.013	0.018	0.667	1.000
		Land surface	3	0.962	0.886	0.018	0.031	0.667	1.000

Table S9 Numerical verification of variable-set subcomposition stability

Omitted variable	Region-stage cases	Maximum absolute residual	Minimum rho	Ordering preserved
GST	10	1.110e-16	1.000	10/10
PRE	10	2.220e-16	1.000	10/10
PRS	10	1.665e-16	1.000	10/10
RHU	10	1.665e-16	1.000	10/10
TEM	10	1.110e-16	1.000	10/10
UV10	10	2.220e-16	1.000	10/10

Comment 2.6

6. The final weight definition requires further clarification and justification. The product of two Softmax-normalized weights does not generally sum to one, so an additional normalization step should be specified if the product is used as the final variable weight. Moreover, the multiplicative structure gives high final weights mainly to variables that are high in both components. The authors should justify this choice and examine whether the selected parameterization schemes are robust to alternative combination rules (e.g., weighted additive formulation).

Response 2.6:

We thank the reviewer for identifying this important issue in the definition and interpretation of the final weights. We agree that the product of two Softmax-normalized component weights does not, by itself, generally sum to one. If this additional normalization step is not stated explicitly, the final variable weights are difficult to reproduce and the method can be misinterpreted. We also agree that the multiplicative combination is not a neutral algebraic detail: it gives high final weights mainly to variables that have high values in both the MRE-based and CV-based components. We therefore revised the manuscript to clarify the formula, justify the non-compensatory product structure, and test an additive alternative as requested.

First, we clarified the final normalization in Sect. 2.3.3. In the actual implementation, the product of the two component weights was already normalized across all variables. The revised manuscript now defines the final weight as

$$\omega_m = \frac{\omega_{m1}\omega_{m2}}{\sum_{j=1}^M \omega_{j1}\omega_{j2}},$$

and explicitly states that this second normalization ensures $\sum_{m=1}^M \omega_m = 1$, as required by the

multivariable evaluation metrics. This revision clarifies an insufficiently described step in the original manuscript rather than changing the underlying calculation.

Second, we expanded the methodological rationale for the multiplicative structure. In the proposed framework, the MRE component identifies variables with large simulation difficulty, whereas the absolute-error CV component identifies variables whose errors respond strongly to the candidate schemes within the evaluated physical-process group. The product rule represents a non-compensatory joint constraint between these two diagnostic signals. A variable receives a high final weight only when it is both difficult to simulate and informative for distinguishing the parameterization choices under evaluation. This is consistent with the purpose of the method: the integrated score should retain the influence of variables that are both high-error and scheme-responsive, rather than being dominated by variables that are high in only one component.

Third, we added an explicit sensitivity test using an equally weighted additive alternative. The additive rule was defined as

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

This additive formulation has a compensatory interpretation, because a high value in one component can offset a lower value in the other. It therefore provides a deliberately more permissive alternative to the proposed multiplicative rule. We applied this additive rule to the stage-wise parameterization screening in both study regions and added the full results to Supplementary Table S10.

The new results show that the additive rule selected the same stage-level cases as the proposed multiplicative rule in all five Southeast China evaluation stages. In Northwest China, it selected the same cases in four of the five stages, with a change only in the cumulus-stage case. These results indicate that the main parameterization-selection outcome was largely stable to this compensatory alternative and was not an artifact of a single algebraic form. At the same time, the Northwest China cumulus-stage difference shows that the multiplicative rule is not equivalent to a simple additive average and can impose a stronger joint-constraint criterion in boundary cases.

For the region where the additive rule produced a different final configuration, we further conducted an independent 2010 validation. In Northwest China, the additive-rule configuration produced cMIEI = 1.312736, whereas the proposed multiplicative-rule configuration produced cMIEI = 1.309856 (reported as 1.313 and 1.310 in the main text). Thus, under the present evaluation objective and dataset, the multiplicative-rule configuration retained a slightly lower integrated cMIEI in the independent validation year. We also bounded this interpretation in the Discussion: these results support the use of the multiplicative combination for the present framework, but do not imply that a product rule would be universally superior for all model-

evaluation contexts.

In summary, we addressed the reviewer's concern by (1) specifying the second normalization step in the final weight formula, (2) explaining the product rule as a non-compensatory joint constraint between simulation difficulty and scheme-dependent response, and (3) adding an equally weighted additive-combination sensitivity test with stage-level and independent-validation results. These revisions clarify the mathematical definition, methodological rationale, and empirical checks of the proposed weighting framework, and further define its scope within the present evaluation framework.

Corresponding revisions:

Section 2.3.3 (page 12, lines 261-265):

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.

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 29, lines 618-619):

The multiplicative structure is central to this interpretation. It was not introduced simply to increase weight contrast, but to represent a non-compensatory joint constraint between the two diagnostic components. The MRE component identifies variables with large simulation difficulty, whereas the absolute-error CV component identifies variables whose errors respond strongly to the candidate schemes within the evaluated physical-process group. Under the proposed product rule, a variable receives a high final weight only when both conditions are satisfied. This design is consistent with the purpose of the framework: the integrated score should emphasize variables that are both difficult to simulate and informative for distinguishing parameterization choices, rather than variables that are high in only one component.

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 29, lines 625-626):

The additive-rule sensitivity test provides a useful contrast to this design. A weighted additive rule treats the MRE and CV components as partially compensatory, so a high value in one component can offset a lower value in the other. It is therefore a more permissive alternative than the proposed multiplicative rule. 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 and was not an artifact of a single algebraic form. At the same time, the Northwest China difference shows that the product rule is not equivalent to a simple additive average and can impose a stronger joint-constraint criterion in boundary cases. In the independent 2010 Northwest China validation, the multiplicative-rule configuration also retained a slightly lower cMIEI than the additive-rule configuration. These results support the use of the multiplicative combination for the present

evaluation objective, while not implying that a product rule would be universally superior for all model-evaluation contexts.

Section 2.3.3 Performance-based multivariable weighting framework (page 13, lines 267-268):

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

Section 2.3.3, added additive-combination sensitivity formula (page 13, lines 269-271):

This additive formulation has a compensatory interpretation, because a large value in one component can partly offset a smaller value in the other component. It was therefore used as a deliberately more permissive alternative to the proposed multiplicative rule, which gives the largest final weights to variables supported by both the MRE and CV components. The selected-scheme consequences of this alternative rule are reported in Table S10.

Section 3.5 Comparison with established statistical weighting methods (page 27, lines 571-580):

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 = 1.313 for the additive-rule configuration, close to but slightly higher than the proposed multiplicative-rule configuration (cMIEI = 1.310). These results indicate that the main scheme-selection outcome was largely stable to this compensatory alternative, while the Northwest cumulus-stage difference shows that the multiplicative rule imposes a distinct joint-constraint interpretation rather than behaving as a simple additive average.

Supporting Information, Table S10:

We added the stage-level sensitivity results for the equally weighted additive combination rule. Table S10 shows that the additive rule selected the same final configuration as the proposed multiplicative rule in Southeast China and changed only the cumulus-stage choice in Northwest China. In the 2010 Northwest China validation, the proposed multiplicative-rule configuration had cMIEI = 1.309856, whereas the additive-rule configuration had cMIEI = 1.312736.

Supporting material cited in this response: Table S10

Table S10 Sensitivity of selected schemes to an equally weighted additive combination rule

(a) Stage-level screening outcomes

Region	Physical-process stage	Proposed MRE $\times$ CV scheme	Proposed stage cMIEI	Additive $\alpha=0.5$ scheme	Additive stage cMIEI	Selection outcome
Northwest China	Microphysics	P3	1.185947	P3	1.170784	Unchanged
	PBL / surface layer	YSU + MM5	1.178968	YSU + MM5	1.164481	Unchanged
	Cumulus	Tiedtke	1.180181	New Tiedtke	1.165333	Changed
	Radiation	GFDL	1.166701	GFDL	1.157858	Unchanged
	Land surface	Noah LSM	1.164356	Noah LSM	1.156148	Unchanged
Southeast China	Microphysics	WDM6	1.036659	WDM6	1.041232	Unchanged
	PBL / surface layer	Shin-Hong + MM5	1.025430	Shin-Hong + MM5	1.032803	Unchanged
	Cumulus	Grell 3	1.031720	Grell 3	1.030902	Unchanged
	Radiation	GFDL	1.011017	GFDL	1.012825	Unchanged
	Land surface	Noah LSM	1.012347	Noah LSM	1.014569	Unchanged

(b) Independent 2010 validation of the additive-rule final configuration

Region	Configuration source	cMIEI	cVSC	cRMSL
Northwest China	Proposed MRE $\times$ CV	1.309856	0.180869	1.137148
	Additive $\alpha=0.5$	1.312736	0.192743	1.232361
Southeast China	Proposed MRE $\times$ CV	1.008	0.508	1.067
	Additive $\alpha=0.5$	1.008	0.508	1.067

Comment 2.7

7. The Introduction is overly lengthy and lacks a clear focus.

Response 2.7:

We thank the reviewer for this helpful comment. We agree that the original Introduction was too broad and did not bring the reader quickly enough to the central methodological problem. We have therefore reorganized and condensed Sect. 1 Introduction.

The revised Introduction now follows a clearer sequence: (1) the need for multivariable model evaluation, (2) WRF parameterization-scheme uncertainty and the computational constraint of

exhaustive scheme screening, (3) the specific limitation of existing multivariable weighting approaches for WRF scheme evaluation, and (4) the contribution and scope of the proposed $MRE \times CV$ weighting framework. We also removed or compressed general background material and made the focus of the study explicit earlier in the section.

This revision was coordinated with our responses to the reviewer's other comments. In particular, the Introduction now avoids the ambiguous term "dynamic weighting", defines the method as a "performance-based multivariable weighting framework", and clarifies that the method changes the evaluation and selection of WRF parameterization-scheme combinations rather than directly modifying the WRF model physics or numerical simulations. The revisions appear in Section 1 Introduction (page 2, lines 39-46 and 52-59).

Corresponding revisions:

Section 1 Introduction (page 2, lines 39-46):

A remaining difficulty is how to evaluate WRF configurations when several meteorological variables are considered together. 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 simulation 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 on subjective expert weighting. These choices can dilute variables with large simulation errors, such as precipitation and wind speed, 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 incorporate both variable-level simulation difficulty and scheme-dependent response.

Section 1 Introduction (page 2, lines 52-59):

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 distinguishing 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.

Comment 2.8

8. The manuscript should more clearly distinguish between model evaluation and model improvement. The proposed weighting method changes how different variables contribute to the evaluation index and may help identify better-performing scheme combinations, but it does not directly improve the WRF simulations themselves. Therefore, statements suggesting that the method “improves the model’s simulations” (line 599) should be revised.

Response 2.8:

We thank the reviewer for pointing out this important distinction. We agree that the original wording could be read as implying that the weighting method directly improves the WRF simulations. This was not our intended meaning. The proposed $\text{MRE} \times \text{CV}$ weighting method does not modify the WRF dynamical core, physical parameterizations, forcing data, or numerical integration. Its role is to change how evaluated variables contribute to the multivariable evaluation index and thereby affect which parameterization-scheme combinations are selected.

To address this concern, we revised the manuscript throughout the Results, Discussion, and Conclusion to distinguish evaluated performance from direct model improvement. In the Results, we replaced statements such as “improvement in model simulation accuracy” with wording that refers to lower integrated evaluation scores, higher R, lower RMSE, or better evaluated agreement of the selected configuration. This keeps the numerical comparison intact while avoiding the implication that the weighting algorithm itself changed the underlying WRF simulations.

We also added an explicit boundary statement in the Discussion. The revised text now states that the framework does not directly modify WRF physical parameterizations; instead, it changes how variables contribute to the integrated evaluation and therefore affects which scheme combinations are selected. We further clarified that lower cMIEI, lower 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. In the Conclusion, we similarly revised the text to state that the method improves parameterization-scheme evaluation and selection, whereas further refinement of WRF physical parameterizations remains necessary for representing complex weather systems and extreme events.

Corresponding revisions:

[Section 3.2 Independent validation of the optimal parameterization-scheme combination for the arid region of Northwest \(page 19, lines 427-434\):](#)

[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.](#)

Section 3.4 Independent validation of the optimal parameterization-scheme combination for the coastal region of Southeast (page 23, lines 492-500):

the configuration selected by the proposed weighting scheme achieved a higher correlation coefficient ($R = 0.502$) compared to the equal-weighting scheme ($R = 0.499$), and a smaller RMSE (1.039 vs. 1.106), indicating stronger evaluated performance under the Taylor and MIEI metrics.

Section 4.1 Evaluation meaning and weight-allocation mechanism (page 28, lines 593-596):

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.

Section 3.1 Optimal parameterization scheme combination for the arid region of Northwest (page 16, lines 364-371):

As the evaluation of physical processes advances, a progressive decrease in MIEI values was observed, indicating that the sequentially selected configurations achieved lower integrated evaluation scores.

Section 5 Conclusions (page 36, lines 845-850):

These remaining biases indicate that the weighting method improves the evaluation and selection of parameterization schemes, whereas further refinement of WRF physical parameterizations remains necessary for representing complex weather systems and extreme events.

Comment 2.9

9. The temporal aggregation used to calculate the evaluation metrics is unclear. The manuscript should specify whether R , RMSE, MIEI, and the Taylor diagrams are based on daily values, monthly means, seasonal means, or climatological averages, because these choices can lead to substantially different conclusions.

Response 2.9:

We agree with the reviewer that the manuscript did not state this point clearly enough. Although the evaluation metrics were calculated from daily station-level paired samples, this was not stated sufficiently clearly in the Methods equations, figure captions, and Supporting Information. This lack of clarity could indeed lead to misunderstanding, because different temporal

aggregation choices can change the numerical meaning of R, RMSE/cRMSE, MIEI/cMIEI, and Taylor-diagram statistics. We have therefore added clarifications in the Methods, the relevant figure and table captions, and the Supporting Information.

Specifically, the R, RMSE/cRMSE, MIEI/cMIEI, and Taylor-diagram statistics used for WRF parameterization-scheme screening and the independent 2010 validation were all calculated from daily station-level June-August model-observation pairs. For each evaluation year, WRF outputs were first extracted at the observation-station locations and then paired with daily station observations, giving 92 summer daily values. The evaluation samples included 49 stations in the arid region of Northwest China and 38 stations in the coastal region of Southeast China, after applying the common missing-value mask. In the revised equation description, the sample index t denotes the valid pooled station-day sample after combining the station and day dimensions. Therefore, these core evaluation metrics were not calculated from monthly means, seasonal means, or climatological averages.

The MIEI and Taylor-diagram calculations used daily station-level data, and the station $\times$ day arrays were flattened into a single sample dimension before the statistical calculations. This is consistent with the revised definition of t in the Methods. We also clarified in Sect. 2.3.3 that the MRE and absolute-error CV weight components were calculated from the same daily station-level model-observation pairs and were then summarized across candidate schemes within each physical-process group.

In addition, we clarified two related points that could otherwise be confused with the metric-calculation scale. First, the 30-year climatological reference in the main text and Supporting Information was used only to identify representative screening years from long-term station observations; it was not used to rank WRF candidate parameterization schemes. Second, the scatter plots in Figs. 4(b-h) and 7(b-h) use monthly station summaries only to visualize the model-observation relationships for individual variables, with monthly means for non-precipitation variables and monthly totals for precipitation. These monthly summaries are not the temporal aggregation used for the Taylor-diagram or MIEI statistics. With these revisions, all weighting-method comparisons correspond to the same daily station-level evaluation sample, so differences among methods mainly reflect the variable-weighting and scheme-ranking framework rather than differences in temporal aggregation.

Corresponding revisions:

[Section 2.3.1 \(page 10, lines 191-196\):](#)

[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, MIEI, and the Taylor diagrams. These metrics were not calculated from monthly means, seasonal means, or climatological averages.

Section 2.3.2 Model performance indicators (page 10, lines 206-208):

In the present WRF evaluation, MIEI and its component statistics were calculated from the same daily station-level June-August samples used in the Taylor diagrams; the 30-year climatological reference was used only for representative-year selection, not for ranking WRF candidate schemes.

Section 2.3.3 Performance-based multivariable weighting framework (page 11, lines 225-227):

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.

Figures 3-7 and Tables 3-4 captions (pages 17-28):

Figs. 3 and 6 now state that all MIEI and Taylor-diagram statistics were calculated from daily station-level June-August values. Figs. 4 and 7 now state that panel (a) uses daily station-level Taylor/MIEI statistics, whereas panels (b-h) use monthly station summaries only for visual comparison. Tables 3 and 4 now state that cMIEI, cVSC, and cRMSL were calculated from daily station-level June-August values.

Supporting Information, Figs. S1-S2 and Tables S2-S3:

The representative-year assessment in Figs. S1-S2 and Tables S2-S3 was based on observed daily station records compared with the 30-year climatological reference. This climatological comparison was used only to select representative screening years. It is distinct from the WRF parameterization-scheme evaluation, for which R, cRMSE/cRMSL, cMIEI, and Taylor diagrams were calculated from daily station-level June-August model-observation pairs in the selected evaluation years.

Comment 2.10

10. The results do not clearly explain how the dynamic weights change the ranking of parameterization schemes compared with equal weighting. The authors should show which variables receive higher weights and how these weights directly affect the selection of the final optimal schemes.

Response 2.10:

Thank you for pointing out this important gap. We agree that the original manuscript compared the final configurations selected by the proposed $\text{MRE} \times \text{CV}$ weighting and equal weight-

ing, but did not sufficiently explain the pathway linking changes in variable weights to changes in scheme ranking and final scheme selection. We therefore revised the Results section and added a new supplementary table to make this pathway explicit.

First, we added a new paragraph in the Results section to clarify that the proposed $\text{MRE} \times \text{CV}$ weighting does not uniformly rescale all candidate schemes. Instead, it changes the integrated MIEI ranking when variables that receive higher weights favor a different candidate scheme. This explanation also clarifies that ranking changes occurred only in selected stages rather than in every stage: among the ten region-process evaluations, the Top-1 scheme changed in three stages relative to equal weighting.

Second, we added Table S12 in the Supporting Information to show the ranking pathway. The table uses a pairwise comparison to connect variable weights with scheme selection. For each rank-reversal stage, A denotes the equal-weight winner and B denotes the $\text{MRE} \times \text{CV}$ -weight winner. We then calculate $D_{\text{eq}} = \text{MIEI}_{\text{eq}}(B) - \text{MIEI}_{\text{eq}}(A)$ and $D_{\text{dyn}} = \text{MIEI}_{\text{dyn}}(B) - \text{MIEI}_{\text{dyn}}(A)$. Because a smaller MIEI indicates better integrated performance, a positive value means that B is worse than A under that weighting scheme, whereas a negative value means that B is better than A. Thus, a change from positive D_{eq} to negative D_{dyn} indicates that the relative ranking of the same two candidate schemes reversed after reweighting. Panel C further calculates the variable-level reweighting effect as $E_m = (\omega_m^{\text{dyn}} - \omega_m^{\text{eq}})\Delta z_m$, where $\Delta z_m = z_{m,B} - z_{m,A}$. A negative value means that the variable reduces the weighted MIEI difference of B relative to A and therefore promotes the $\text{MRE} \times \text{CV}$ -weight winner, whereas a positive value indicates the opposite effect.

Panel A reports, for each region and physical-process stage, the highest-weight variables, the equal-weight Top-1 scheme, the $\text{MRE} \times \text{CV}$ -weight Top-1 scheme, whether the Top-1 scheme changed, and the scheme carried forward to the next stage. The table shows that UV10 and PRE generally received the highest weights in Northwest China, for example $\text{UV10} = 0.435$ and $\text{PRE} = 0.233$ in the PBL/surface-layer stage, and $\text{UV10} = 0.422$ and $\text{PRE} = 0.243$ in the radiation stage. In the Southeast China cumulus stage, $\text{PRE} = 0.366$ and $\text{UV10} = 0.321$ were the two highest-weight variables. The performance of candidate schemes for these high-weight variables can therefore affect the stage-wise integrated ranking.

Third, Panels B and C of Table S12 provide pairwise score-difference decompositions for the three rank-reversal cases. In the Northwest PBL/surface-layer stage, equal weighting selected MYJ + MYJ, whereas $\text{MRE} \times \text{CV}$ weighting selected YSU + MM5; the pairwise difference changed from $D_{\text{eq}} = 0.0047$ under equal weighting to $D_{\text{dyn}} = -0.0056$ under $\text{MRE} \times \text{CV}$ weighting. The variable-level terms show that the higher UV10 weight was the main factor promoting YSU + MM5 ($E_m = -0.0097$), with a smaller contribution from GST ($E_m = -0.0015$), whereas PRE opposed this selection ($E_m = 0.0021$). In the Northwest radiation stage, equal weighting selected RRTMG, whereas $\text{MRE} \times \text{CV}$ weighting selected GFDL; the pairwise difference changed from $D_{\text{eq}} = 0.0074$ to $D_{\text{dyn}} = -0.0119$, mainly because UV10 promoted

GFDL ($E_m = -0.0188$). In the Southeast cumulus stage, equal weighting selected New Tiedtke, whereas MRE $\times$ CV weighting selected Grell 3; the pairwise difference changed from $D_{eq} = 0.0013$ to $D_{dyn} = -0.0059$, mainly because PRE promoted Grell 3 ($E_m = -0.0043$).

These additions provide a more specific account of the ranking pathway in the rank-reversal stages. Equal weighting allows all variables to contribute with the same share, whereas the proposed MRE $\times$ CV weighting gives greater influence to variables with larger simulation difficulty and stronger scheme-dependent response. When those high-weight variables favor a different candidate scheme, the Top-1 ranking can reverse; when they do not, the ranking remains unchanged. The revised manuscript therefore provides a more explicit account of which variables received higher weights and how, within the MIEI ranking framework used in this study, these weights affected the final optimal scheme selection.

Corresponding revisions:

[Section 3.4 Independent validation of the optimal parameterization-scheme combination for the coastal region of Southeast \(page 24, lines 512-522\):](#)

[To further clarify how the proposed MRE \$\times\$ CV weights changed the scheme-selection pathway relative to equal weighting, Table S12 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 MIEI 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.](#)

[Supporting Information, Table S12:](#)

[Pathway linking MRE \$\times\$ CV variable weights, rank reversals, and final parameterization-scheme selection.](#)

Supporting material cited in this response: Table S12

Table S12. Pathway linking $\text{MRE} \times \text{CV}$ variable weights, rank reversals, and final parameterization-scheme selection

Panel A. Stage-by-stage pathway from variable weights to selected schemes

Region	Physical-process stage	Highest-weight variables	Equal-weight Top-1 scheme	$\text{MRE} \times \text{CV}$ -weight Top-1 scheme	Top-1 changed
Northwest China	Microphysics	UV10 (0.443) PRE (0.224)	P3	P3	No
	PBL/surface layer	UV10 (0.435) PRE (0.233)	MYJ + MYJ	YSU + MM5	Yes
	Cumulus	UV10 (0.428) PRE (0.275)	Tiedtke	Tiedtke	No
	Radiation	UV10 (0.422) PRE (0.243)	RRTMG	GFDL	Yes
	Land surface	UV10 (0.428) PRE (0.224)	Noah LSM	Noah LSM	No
Southeast China	Microphysics	UV10 (0.366) PRE (0.293)	WDM6	WDM6	No
	PBL/surface layer	UV10 (0.385) PRE (0.257)	Shin-Hong + MM5	Shin-Hong + MM5	No
	Cumulus	PRE (0.366) UV10 (0.321)	New Tiedtke	Grell 3	Yes
	Radiation	PRE (0.325) UV10 (0.317)	GFDL	GFDL	No
	Land surface	PRE (0.345) UV10 (0.319)	Noah LSM	Noah LSM	No

Panel B. Pairwise score-difference decomposition for rank-reversal cases

Case ID	Region	Physical- process stage	A: equal-weight winner	B: MRE × CV-weight winner	$\Delta I_{\text{eq}} =$ $I_{\text{eq},B} - I_{\text{eq},A}$	$D_{\text{dyn}} =$ $\text{MIEI}_{\text{dyn},B} -$ $\text{MIEI}_{\text{dyn},A}$	Total reweighting effect = $\Delta I_{\text{MRE} \times \text{CV}} -$ ΔI_{eq}	Main variables promoting the MRE × CV-weight winner after reweighting	Main variables opposing the MRE × CV-weight winner after reweighting
B1	Northwest China, 2007	PBL/surface layer	MYJ + MYJ	YSU + MM5	0.0047	-0.0056	-0.0103	UV10 (-0.0097) GST (-0.0015)	PRE (0.0021)
B2	Northwest China, 2007	Radiation	RRTMG	GFDL	0.0074	-0.0119	-0.0192	UV10 (-0.0188) RHU (-0.0027)	PRE (0.0047); GST (0.0008)
B3	Southeast China, 2009	Cumulus	New Tiedtke	Grell 3	0.0013	-0.0059	-0.0073	PRE (-0.0043) RHU (-0.0017)	PRS (0.0011)

Panel C. Variable-level contribution terms for rank-reversal cases

Case ID	Variable	w_m^{eq}	$w_m^{\text{MRE} \times \text{CV}}$	$z_{A,m}$	$z_{B,m}$	$\Delta z_m = z_{B,m} - z_{A,m}$	$w_m^{\text{eq}} \Delta z_m$	$w_m^{\text{MRE} \times \text{CV}} \Delta z_m$	$E_m = (w_m^{\text{MRE} \times \text{CV}} - w_m^{\text{eq}}) \Delta z_m$	Effect on MRE $\times$ CV-weight winner
B1	GST	0.1667	0.0912	0.8630	0.8825	0.0196	0.0033	0.0018	-0.0015	promotes
	PRE	0.1667	0.2335	1.0900	1.1220	0.0320	0.0053	0.0075	0.0021	opposes
	PRS	0.1667	0.0643	0.1982	0.2074	0.0092	0.0015	0.0006	-0.0009	promotes
	RHU	0.1667	0.0951	0.7529	0.7546	0.0017	0.0003	0.0002	-0.0001	promotes
	TEM	0.1667	0.0814	0.5332	0.5350	0.0019	0.0003	0.0002	-0.0002	promotes
	UV10	0.1667	0.4347	1.2065	1.1703	-0.0362	-0.0060	-0.0157	-0.0097	promotes
B2	GST	0.1667	0.0983	0.8630	0.8506	-0.0124	-0.0021	-0.0012	0.0008	opposes
	PRE	0.1667	0.2428	1.0900	1.1517	0.0617	0.0103	0.0150	0.0047	opposes
	PRS	0.1667	0.0630	0.1982	0.2223	0.0241	0.0040	0.0015	-0.0025	promotes
	RHU	0.1667	0.0900	0.7529	0.7879	0.0350	0.0058	0.0032	-0.0027	promotes
	TEM	0.1667	0.0837	0.5332	0.5425	0.0094	0.0016	0.0008	-0.0008	promotes
	UV10	0.1667	0.4222	1.2065	1.1329	-0.0736	-0.0123	-0.0311	-0.0188	promotes
B3	GST	0.1667	0.0797	0.9395	0.9549	0.0154	0.0026	0.0012	-0.0013	promotes
	PRE	0.1667	0.3664	1.0679	1.0465	-0.0214	-0.0036	-0.0078	-0.0043	promotes
	PRS	0.1667	0.0688	0.7300	0.7183	-0.0116	-0.0019	-0.0008	0.0011	opposes
	RHU	0.1667	0.0837	1.0185	1.0389	0.0204	0.0034	0.0017	-0.0017	promotes
	TEM	0.1667	0.0805	0.7939	0.8018	0.0079	0.0013	0.0006	-0.0007	promotes
	UV10	0.1667	0.3209	0.9914	0.9887	-0.0028	-0.0005	-0.0009	-0.0004	promotes

Comment 2.11

11. The Taylor diagrams show that many parameterization schemes are clustered closely together, and the improvement of the optimal scheme appears visually small. The authors should provide statistical significance tests, confidence intervals, bootstrap resampling, or cross-validation analyses to show whether the differences between the selected schemes exceed the uncertainty arising from temporal variability, station sampling, and observational errors.

Response 2.11:

We thank the reviewer for this helpful suggestion. We agree that Taylor diagrams provide a useful visual summary of correlation, standard deviation, and error structure, but that closely clustered schemes require an additional quantitative stability assessment. We therefore added a bootstrap-based uncertainty analysis for the selected stage-level parameterization schemes and incorporated the results into the Methods, Results, Discussion, Conclusions, and Supporting Information.

We distinguished two levels of inference. For the study region, screening year, summer period, station network, and observational dataset used in this study, the station distribution and temporal variability are part of the evaluation target, not errors to be removed. The scheme with the lowest full-sample cMIEI is therefore the best-performing configuration for the evaluated station-time sample. The reviewer's uncertainty question mainly concerns a second level of inference: whether this ranking remains stable when the finite station and time samples vary. We therefore frame the added analysis as a test of finite-sample stability in scheme ranking, not as treating station differences or temporal variability as invalid noise.

The added analysis contains two complementary resampling components. First, we used station bootstrap resampling to evaluate the sensitivity of scheme ranking to station sampling. Each bootstrap sample retained the original number of stations, 49 stations in Northwest China and 38 stations in Southeast China, sampled with replacement. Second, we used a 14 d moving-block bootstrap to evaluate sensitivity to intra-seasonal temporal sampling while retaining the continuity of weather-scale variability. For each region-stage comparison, we generated 500 station-bootstrap samples and 500 time-block-bootstrap samples. For each resample, the variable weights, cMIEI values, and candidate-scheme rankings were recalculated. Thus, all candidate schemes were compared within the same resampled observational constraint, weighting formula, and evaluation metric.

We added Figure S8 and Table S13 to the Supporting Information. Figure S8 reports the 95% bootstrap intervals of ΔcMIEI between the selected scheme and nearby competitors. Table S13 reports the proportions of bootstrap samples in which the selected scheme remained in the top 1, top 2, and top 3. The nearest-competitor comparison showed stable separation in a subset of the region-stage-resampling cases, while the rank-retention analysis indicated that the selected

schemes generally remained in the leading group: the median top-1, top-2, and top-3 retention rates were 0.527, 0.857, and 0.959, respectively.

It should also be noted that, with the continuing development of WRF physical parameterization schemes, some schemes share related physical assumptions or are designed to represent similar process targets. Therefore, closely clustered performance among several candidate schemes in a given region, season, and evaluation-variable setting is a reasonable outcome. This does not diminish the practical value of parameterization-scheme screening. In practical WRF applications, researchers often need to identify one or several competitive configurations that match the target region, evaluated variables, and application objective within a limited number of simulations. The proposed $\text{MRE} \times \text{CV}$ weighting framework is not intended to demonstrate that one scheme is uniquely optimal under all conditions, but to help identify candidate configurations that have favorable overall performance and application potential under a prescribed multivariable evaluation objective in a suitable application context. When several configurations show only small differences, subsequent application-oriented studies may further compare these competitive candidates under the specific scientific question of interest.

We revised the manuscript accordingly. Rather than relying only on the visual separation in the Taylor diagrams, the revised text now presents the selected schemes as configurations that performed well in the full-sample evaluation and generally remained within the leading group under station and temporal resampling. This strengthens the methodological argument of the paper: the proposed $\text{MRE} \times \text{CV}$ framework not only changes the selection pathway according to limiting variables and scheme-dependent responses, but also identifies parameterization configurations with good ranking stability in the evaluated station-time sample.

Regarding observational errors, the study uses quality-controlled daily station observations as the common reference dataset. In every bootstrap sample, all candidate schemes are evaluated against the same resampled observations. The added analysis therefore quantifies uncertainty associated with station sampling and intra-seasonal temporal sampling in a paired comparison framework. We did not impose an additional observational-error perturbation model because the present dataset does not provide station- and variable-specific instrument-error distributions or error covariance information. This avoids introducing unverified assumptions while still directly addressing the sampling components raised by the reviewer.

Corresponding revisions:

[Section 2.3.3 Performance-based multivariable weighting framework \(page 13, lines 285-289\):](#)

[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% boot-](#)

strap 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 S13).

Section 3.1 Optimal parameterization scheme combination for the arid region of Northwest (page 17, lines 386-393):

The bootstrap ranking analysis supported this interpretation for Northwest China (Fig. S8 and Table S13). 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% ΔcMIEI 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.

Section 3.3 Optimal parameterization-scheme combination for the coastal region of Southeast (page 21, lines 476-479):

For Southeast China, the selected schemes also generally remained in the leading group under resampling (Fig. S8 and Table S13). The median top-three retention rates were 0.822 in station-bootstrap samples and 0.944 in time-block-bootstrap samples. 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.

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 33, lines 751-752):

The scheme-ranking analysis complements the weight-stability results by evaluating whether the selected stage-level 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.

Supporting Information, Fig. S8 and Table S13:

We added Figure S8. Scheme-ranking uncertainty of the selected stage-level parameterization schemes and Table S13. Rank retention of the full-sample selected stage-level schemes under station and temporal resampling.

Supporting material cited in this response: Fig. S8 and Table S13

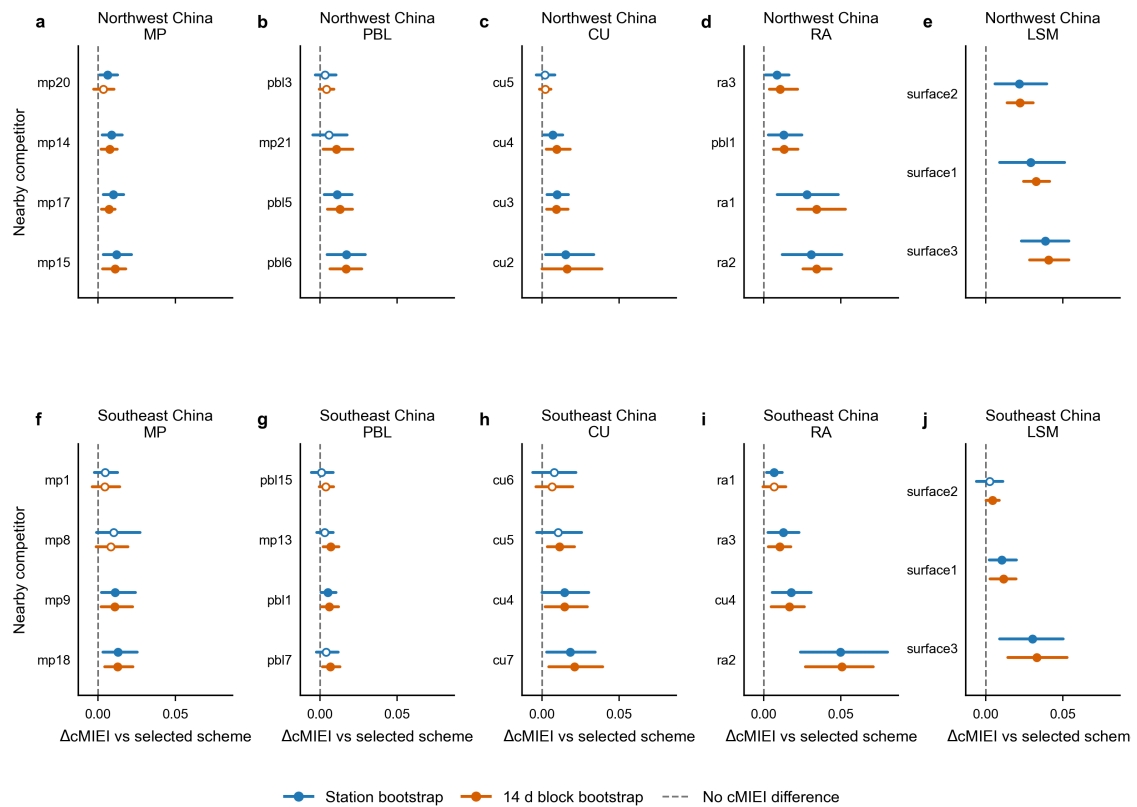


Figure S8. Bootstrap uncertainty in the cMIEI advantage of the full-sample selected stage-level schemes over nearby competitors. Blue circles indicate station-bootstrap samples, and orange squares indicate 14 d moving-block-bootstrap samples. Horizontal bars show 95% bootstrap intervals of $\Delta cMIEI$. Positive values indicate that the selected scheme had a lower cMIEI than the competing scheme. Intervals entirely to the right of zero indicate stable separation under the corresponding resampling analysis.

Table S13 Rank retention of the full-sample selected stage-level schemes under station and temporal resampling

Region	Evaluation stage	Full-sample selected scheme	Perturbation	Top-1 retention	Top-2 retention	Top-3 retention
Northwest China	Cumulus	Tiedtke	Station bootstrap	0.418	0.768	0.918
			14 d block bootstrap	0.518	0.818	0.942
	Land surface	Noah LSM	Station bootstrap	0.938	1.000	1.000
			14 d block bootstrap	1.000	1.000	1.000
	Microphysics	P3	Station bootstrap	0.678	0.906	0.974
			14 d block bootstrap	0.536	0.822	0.870
	PBL / surface layer	YSU + MM5	Station bootstrap	0.432	0.750	0.896
			14 d block bootstrap	0.666	0.892	0.984
	Radiation	GFDL + GFDL	Station bootstrap	0.792	0.926	0.994
			14 d block bootstrap	0.998	1.000	1.000
Southeast China	Cumulus	Grell 3	Station bootstrap	0.500	0.694	0.822
			14 d block bootstrap	0.504	0.804	0.944
	Land surface	Noah LSM	Station bootstrap	0.406	0.896	0.998
			14 d block bootstrap	0.712	0.992	0.998
	Microphysics	WDM6	Station bootstrap	0.444	0.658	0.800
			14 d block bootstrap	0.436	0.660	0.808
	PBL / surface layer	Shin-Hong + MM5	Station bootstrap	0.196	0.404	0.564
			14 d block bootstrap	0.384	0.644	0.840
	Radiation	GFDL + GFDL	Station bootstrap	0.736	0.934	0.990
			14 d block bootstrap	0.718	0.976	1.000

References cited in the responses to Reviewer 2

Baker, N. C. and Taylor, P. C.: A framework for evaluating climate model performance met-

rics, *J. Clim.*, 29, 1773-1782, doi:10.1175/JCLI-D-15-0114.1, 2016.

Dai, D., Chen, L., Ma, Z., and Xu, Z.: Evaluation of the WRF physics ensemble using a multivariable integrated evaluation approach over 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Response to Reviewer 3

General Comments

In general, I consider that the manuscript addresses a relevant problem for the WRF modeling community, since the selection of physical parameterization schemes remains one of the main sources of uncertainty in regional atmospheric simulations. The proposed weighting strategy based on the performance of multiple variables is interesting and has the potential to contribute to a more comprehensive evaluation of model configurations. However, I believe that some aspects could be strengthened to increase the scientific value of the study.

Response to General Comments:

We sincerely thank the reviewer for the positive and constructive overall assessment of our manuscript. We appreciate the reviewer's recognition that WRF physical parameterization-scheme selection is an important source of uncertainty in regional atmospheric simulations, and that the proposed performance-based multivariable weighting strategy has the potential to support more comprehensive evaluation of model configurations.

In response to the reviewer's comments, we strengthened the manuscript in three main aspects. First, we expanded the physical interpretation of the selected parameterization-scheme combinations by linking the statistical selections to the dominant regional processes in the arid Northwest and humid coastal Southeast China. Second, we clarified the scope of uncertainty addressed by the proposed framework, emphasizing that this study evaluates uncertainty associated with WRF physical parameterization-scheme selection under a fixed model configuration, rather than all sources of WRF simulation uncertainty. Third, we added a discussion of the practical meteorological meaning of the evaluated differences by connecting changes in integrated scores to variable-level RMSE changes in physically interpretable units.

These revisions were made to ensure that the methodological contribution is not presented only as a statistical ranking procedure, but is also connected to regional physical consistency, practical meteorological interpretation, and a clearly defined uncertainty scope. The above provides our overall response to the reviewer's General Comments. Detailed point-by-point responses are provided below.

In the point-by-point responses below, our responses are shown in regular black font, [whereas the corresponding revisions made in the manuscript are shown in blue regular font.](#)

Comment 3.1

The manuscript would benefit from a deeper discussion of the physical interpretation of the selected parameterization schemes. Although the methodology successfully identifies the configurations that provide the best statistical performance, the discussion remains focused mainly on the evaluation metrics. It would be valuable to explain why the selected schemes may be physically suitable for the climatic and atmospheric characteristics of the analyzed regions. Such discussion would help readers understand whether the methodology is identifying parameterizations that are not only statistically optimal but also physically consistent with the dominant atmospheric processes.

Response 3.1:

Thank you for this helpful comment. We agree that readers need to understand whether the statistically selected parameterization combinations are also physically consistent with the dominant atmospheric processes in the two study regions.

In response, we added a new Sect. 4.2, “Physical interpretation and practical meteorological relevance”. In this section, we discuss the two selected optimal configurations separately. For the arid Northwest domain, we explain why the P3, YSU, Tiedtke, GFDL, and Noah LSM combination is physically plausible in relation to intermittent precipitation, complex terrain, boundary-layer exchange, radiation, and land-surface energy-water exchanges. For the humid coastal Southeast domain, we explain why the WDM6, Shin-Hong-MM5, Grell 3, GFDL-GFDL, and Noah LSM combination is consistent with monsoon moisture transport, moist convection, cloud-precipitation processes, and coastal boundary-layer dynamics.

We also revised the tone of the discussion to avoid claiming a causal mechanism that was not directly diagnosed. The revised text therefore states that the selected schemes are physically consistent with the dominant regional processes, while acknowledging that water-vapor budget analysis, hydrometeor diagnostics, turbulent kinetic energy diagnostics, PBL-height evaluation, and surface energy-balance decomposition would be needed for formal process attribution.

Corresponding revisions:

[Section 4.2 Physical interpretation and practical meteorological relevance \(page 31, lines 673-676\):](#)

[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 particularly sensitive to cloud microphysics, boundary-layer mixing, radiation, and surface energy-water exchanges. The selection of the P3 microphysics scheme is physically plausible because P3 predicts bulk ice particle properties and explicitly represents the mass and volume of rimed ice, allowing a more flexible treatment of mixed-phase hydrometeors](#)

than schemes with fixed ice categories (Morrison and Milbrandt, 2015). This is relevant for precipitation simulation over arid and topographically complex regions, where precipitation events are intermittent and strongly affected by cloud microphysical uncertainty. The selection of the YSU PBL scheme is also consistent with the importance of near-surface wind and temperature in this region, because YSU uses a non-local K-profile approach with an explicit entrainment layer to represent turbulent exchange in the unstable mixed layer (Hong et al., 2006). The Tiedtke cumulus scheme, a mass-flux scheme with shallow convection and momentum transport, provides a physically meaningful representation of sub-grid convective transport under conditions where convection is not fully resolved (Tiedtke, 1989; Zhang et al., 2011). Together with the GFDL radiation scheme, which accounts for radiative transfer with carbon dioxide, ozone, and cloud microphysical effects (Fels and Schwarzkopf, 1981), and the Noah LSM, which represents soil temperature, soil moisture, snow, and surface fluxes in the land-atmosphere system (Chen and Dudhia, 2001), this combination is consistent with the dominant controls on wind, temperature, and intermittent precipitation in the arid Northwest domain.

Section 4.2 Physical interpretation and practical meteorological relevance (page 31, lines 691-694):

In the humid coastal region of Southeast China, the dominant simulation challenges are more closely related to monsoon moisture transport, moist convection, cloud-precipitation processes, and the coastal boundary layer. The selection of WDM6 is physically interpretable because this scheme extends WSM6 by including double-moment treatment of rain and cloud water as well as prognostic cloud condensation nuclei for warm-rain processes, which is relevant for warm and moist precipitation environments (Lim and Hong, 2010). The selected Shin-Hong PBL scheme includes scale dependency for vertical transport in the convective boundary layer while retaining YSU-type treatment for stable boundary-layer and free-atmospheric mixing (Shin and Hong, 2015). This feature is consistent with the need to represent vertical exchange of heat, moisture, and momentum in coastal regions where sea-land thermal contrast and moist boundary-layer processes affect near-surface wind and humidity. The Grell 3 cumulus scheme is an improved Grell-Devenyi type mass-flux approach and can represent sub-grid convective effects through ensemble-based convective assumptions (Grell, 1993; Grell and Devenyi, 2002). Its selection in the Southeast domain is therefore consistent with the higher weight assigned to precipitation in the cumulus evaluation and with previous studies showing that convection and microphysics parameterizations strongly influence precipitation and cloud properties over Asia and East Asian summer monsoon rainfall (Liu et al., 2018; Yang et al., 2015).

Section 4.2 Physical interpretation and practical meteorological relevance (page 32, lines 707-708):

These physical interpretations should be treated as mechanism-informed explanations of the statistical selection results, not as direct process-attribution evidence.

Comment 3.2

The uncertainty analyzed in this study is associated with the selection of physical parameterization schemes. However, parameterization choice represents only one component of the overall uncertainty present in WRF simulations. Other factors, such as spatial resolution, domain configuration, land-use representation, topography, initial and boundary conditions, and spin-up time may also influence model performance. A brief discussion clarifying the specific scope of the proposed methodology and its relationship to other sources of uncertainty would strengthen the interpretation of the results.

Response 3.2:

We thank the reviewer for pointing out this important distinction. We agree that physical parameterization choice is only one component of the broader uncertainty in WRF simulations, and that the original manuscript needed to define the scope of the proposed framework more explicitly.

In response, we revised the manuscript in three connected places. First, in Sect. 2.2 and Table 1, we clarified the simulation and evaluation periods. The simulations covered May 1 to August 31, with May 1 to May 31 used as the model spin-up period and only June-August used for evaluating the candidate parameterization-scheme combinations. This clarification makes the spin-up design explicit rather than leaving it implicit in the text.

Second, we added a scope statement in Sect. 4. The revised Discussion now states that the uncertainty addressed in this study is the uncertainty associated with WRF physical parameterization-scheme selection under a fixed model configuration. In the present experimental design, all candidate simulations share the same WRF dynamical core, horizontal resolution, domain configuration, vertical setup, initial and boundary conditions, land-use dataset, spin-up design, evaluation variables, and observational reference. Therefore, the reported differences should be interpreted as differences in the evaluated multivariable performance of alternative parameterization-scheme configurations and weighting choices within this controlled framework.

Third, we explicitly acknowledge that other sources of WRF uncertainty, including horizontal and vertical resolution, domain extent and nesting strategy, terrain and land-use representation, initial and boundary forcing, and spin-up length, can also influence model performance. These factors were not independently perturbed or attributed in the present study. We therefore revised the manuscript to state that the proposed $\text{MRE} \times \text{CV}$ weighting framework is a reproducible evaluation tool for parameterization-scheme screening, rather than a complete uncertainty-decomposition framework for WRF simulations.

This revision aligns the methodological claim more closely with the actual experimental design. The role of the method is to improve the transparency of multivariable parameterization-

scheme evaluation under a fixed WRF setup and to provide additional information for variable-level diagnostic interpretation, rather than to replace a full uncertainty attribution across resolution, domains, forcing data, land-use representation, terrain, and spin-up choices. This clarification defines the interpretation range of the present results and points to future applications that combine the weighting framework with broader WRF sensitivity experiments.

Corresponding revisions:

Table 1 (page 5, lines 118-125):

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.

Section 4.3 Robustness and uncertainty boundaries of the screening framework (page 34, lines 782-789):

The uncertainty addressed in this study should therefore be interpreted as uncertainty associated with WRF physical parameterization-scheme selection under a fixed model configuration. In the present experiments, all candidate simulations shared the same WRF dynamical core, horizontal resolution, domain configuration, vertical setup, initial and boundary conditions, land-use dataset, spin-up design, evaluation variables, and observational reference. This controlled design allows the proposed framework to diagnose how alternative physical parameterization schemes and variable-weighting choices affect the evaluated multivariable performance. It also keeps the comparison focused on the parameterization-screening problem that the method is designed to solve. Other sources of WRF uncertainty, including horizontal and vertical resolution, domain extent and nesting strategy, terrain and land-use representation, initial and boundary forcing, and spin-up length, can also influence model performance. These factors were not independently perturbed or attributed in the present study. The proposed $\text{MRE} \times \text{CV}$ weighting framework should therefore be viewed as a reproducible evaluation tool for parameterization-scheme screening, rather than as a complete uncertainty-decomposition framework for WRF simulations. Future applications could combine this weighting framework with multi-resolution, multi-domain, land-use, forcing, or spin-up sensitivity experiments to assess how parameterization-scheme uncertainty interacts with these additional uncertainty sources.

Table 1 (page 5, lines 104-105):

May 1st-Aug 31st, 2007 / May 1st-Aug 31st, 2009; Spin-up period: May 1st-May 31st; Evaluation period: June 1st-Aug 31st.

Comment 3.3

The manuscript demonstrates differences among the evaluated parameterization combinations; however, it would be useful to discuss the practical implications of these differences from a meteorological perspective. In several cases, the statistical differences between competing configurations appear relatively small. Therefore, it would be interesting to discuss whether these improvements translate into a substantially better representation of the atmospheric processes of interest or whether they mainly reflect differences in the evaluation metrics. This would help readers better assess the practical significance of the proposed methodology.

Response 3.3:

We thank the reviewer for this constructive suggestion. We agree that the manuscript should explain the practical meteorological meaning of the differences among the selected parameterization configurations, rather than presenting the differences only as changes in integrated scores.

In response, we revised Sect. 4.2 to connect the integrated evaluation results with the meteorological variables and physical-process groups that drive the scheme-selection pathway. The revised Discussion now clarifies that differences between competing configurations are not only numerical changes in the evaluation metric. They also indicate which parameterization choices better reproduce meteorological variables tied to the atmospheric processes of interest.

Specifically, the $\text{MRE} \times \text{CV}$ -selected configuration reduced RMSE for temperature-related variables in Northwest China (TEM by $0.98\text{ }^{\circ}\text{C}$ and GST by $2.47\text{ }^{\circ}\text{C}$) and for precipitation in Southeast China (PRE by 1.38 mm d^{-1}). Because RMSE retains the unit of the evaluated variable, these reductions can be interpreted in the physical units of the corresponding meteorological variables. These variable-unit diagnostics were used only to interpret the practical meteorological meaning of the selected configurations; they were not introduced as an additional scheme-ranking criterion. In other words, the scheme rankings remained based on the cMIEI-based evaluation framework, whereas the variable-unit diagnostics provided an interpretive bridge between the integrated scores and meteorological variables with physical units.

We further expanded the Discussion to translate these variable-level improvements into meteorological implications. Lower temperature-related errors imply a more reliable representation of near-surface thermal conditions and land-atmosphere heat exchange, which is relevant for regional climate simulation and heat-related extreme-event assessment in arid Northwest China. Lower precipitation errors improve the representation of water input and rainfall-forming processes, which is important for hydroclimate simulation and heavy-precipitation evaluation in humid Southeast China. Near-surface wind is also emphasized because it constrains boundary-layer mixing and moisture and heat transport.

We therefore revised the manuscript to state that the proposed weighting framework has practical value for selecting physical-process parameterizations that better represent target meteorological processes and their application-relevant variables. We also retained the appropriate scientific boundary: formal attribution of the underlying mechanisms would still require process-budget diagnostics such as water-vapor budget analysis, hydrometeor diagnostics, turbulent kinetic energy diagnostics, PBL-height evaluation, or surface energy-balance decomposition.

Corresponding revisions:

Section 4.2 Physical interpretation and practical meteorological relevance (page 32, lines 714-715):

The practical meteorological implication of the evaluated differences becomes clearer when the integrated scores are linked to the variables and physical processes that drive the scheme-selection pathway. Differences between competing configurations are not only numerical changes in the evaluation metric; they indicate which parameterization choices better reproduce meteorological variables tied to the atmospheric processes of interest. For example, the proposed $\text{MRE} \times \text{CV}$ configuration reduced RMSE for temperature-related variables in Northwest China (TEM by 0.98 °C and GST by 2.47 °C) and for precipitation in Southeast China (PRE by 1.38 mm d⁻¹). These variable-level improvements translate into practical meteorological meaning because the affected variables control different parts of the regional climate system. Lower temperature-related errors imply a more reliable representation of near-surface thermal conditions and land-atmosphere heat exchange, which is relevant for regional climate simulation and heat-related extreme-event assessment in arid Northwest China. Lower precipitation errors improve the representation of water input and rainfall-forming processes, which is important for hydroclimate simulation and heavy-precipitation evaluation in humid Southeast China. Near-surface wind is also a key constraint on boundary-layer mixing and moisture and heat transport. The proposed weighting framework therefore has practical value for selecting physical-process parameterizations that better represent target meteorological processes and their application-relevant variables, while process-budget diagnostics would still be needed for formal attribution of the underlying mechanisms.

References cited in the responses to Reviewer 3

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

Grell, G. A.: Prognostic evaluation of assumptions used by cumulus parameterizations, *Mon. Weather Rev.*, 121, 764-787, 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.

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.

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.