

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

Response to the Editor

Editor comment

Public justification (visible to the public if the article is accepted and published):

Dear authors,

Thank you for your thoroughly revised version. Because of the quite substantial changes, I asked one of the reviewers for a second opinion on the new version of the manuscript. The reviewer raised a few more points (please see supplement .pdf) and I would like to ask you to consider the reviewer's points before the paper can be accepted for publication. As recommended by the reviewer, I'll classify this as 'major revisions' even though I would not expect that very substantial changes are required.

Thanks again,

Axel Lauer

(handling topical editor)

Response:

We sincerely thank the editor for continuing to coordinate the review of our manuscript, for acknowledging our previous revisions, and for giving us the opportunity to further improve the manuscript. We are also grateful to the reviewer for the renewed careful assessment and constructive suggestions. Following your recommendation, we have considered and responded to each comment and revised the manuscript and Supporting Information accordingly, focusing on the mathematical interpretation of the weighting rule, the evidence for cross-method comparisons, and the clarity of the results.

First, we have clarified the compensatory mathematical nature of the $MRE \times CV$ combination rule and revised the associated interpretation and wording. We have also added an MRE-only baseline to Fig. 8 and reported the stagewise weight changes and selected schemes in Supporting Information Table S11 to directly examine the information provided by CV beyond MRE alone.

Second, we have added a common variable-level comparison of the final configurations selected by all five methods. Supporting Information Table S15 reports unweighted RMSE, MAE, mean bias, and correlation for all seven evaluated variables, using common valid daily station samples within each region-year and covering both the screening years and the 2010 validation cases. The corresponding Results and Discussion now use these comparisons to explain variable-level improvements and trade-offs and identify the reference configurations for the reported error differences.

In addition, we have clarified the use of common reference weights across configurations in Tables 3 and 4 and added a complete cross-evaluation matrix under the five weighting rules in Supporting Information Table S16. Table S17 further documents the complete stepwise se-

lection pathways for Entropy, CV, and CRITIC. These additions allow clearer comparisons of the selection processes, final configurations, and their performance under different evaluation objectives.

Finally, we have harmonized variable symbols, wind-component definitions, and metric terminology; clarified temporal aggregation and display transformations in the figures; revised wording concerning statistical significance; and strengthened the links between the physical interpretation of the parameterization schemes and the variable-level results. We have also explained the conditional nature of configuration selection under the prescribed screening order.

We hope that these revisions satisfactorily address the concerns raised in this round of review and make the methodological description and supporting evidence clearer and more complete. Our detailed responses and the corresponding manuscript revisions are provided below, following the original order of the reviewer's comments.

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.](#)

Response to the Reviewer

General Comments

The manuscript presents a potentially useful framework for multivariable WRF parameterization screening. And the authors have revised the manuscript following the comments. However, several methodological issues currently limit the robustness and comparability of the results. In particular, the mathematical interpretation of the proposed weighting rule and the comparison among weighting methods require substantial clarification. I therefore recommend major revision. Additional issues concerning terminology, figures, and presentation are listed as minor comments.

Response to General Comments:

We sincerely thank the reviewer for the continued careful assessment of our manuscript and for recognizing the potential usefulness of the proposed framework. We appreciate the reviewer's further comments on the mathematical interpretation, methodological comparability, and presentation of the results. We have treated this overall assessment as a guide for the present revision, with particular attention to making the combination rule transparent and providing direct, comparable evidence of how the weighting methods affect variable-level performance and parameterization-scheme selection.

First, we have clarified the mathematical interpretation of the $MRE \times CV$ rule and revised the associated claims throughout the manuscript. The normalized product is now explicitly described as compensatory, with its equivalence to applying Softmax to the sum of the MRE and CV diagnostic scores explained in the response and reflected in the revised interpretation. We also distinguish this rule from the additive alternative that averages the two separately normalized component-weight distributions. To examine the contribution of CV beyond MRE alone, we have added an MRE-only baseline to Fig. 8 and reported the stagewise absolute weight changes and selected schemes in Supporting Information Table S11. This comparison clarifies how CV supplements the dominant MRE pattern with information about stage-dependent differences in error response.

Second, we have expanded the final-configuration comparisons to address the reviewer's concern about meteorological improvements and potential losses in other variables. Table S15 now reports unweighted RMSE, MAE, mean bias, and Pearson correlation for all seven evaluated variables and all five methods in both screening years and the two regional 2010 validation cases, using common valid daily station samples within each region-year. The corresponding Results and Discussion have been revised to relate the integrated evaluation to the observed variable-level improvements and trade-offs, with explicit reference configurations for the reported error differences.

We have also clarified that Tables 3 and 4 use one common reference-weight vector for all configurations within each region-year. Supporting Information Table S16 provides a complete cross-evaluation matrix in which every final configuration is evaluated under each of the five weighting rules, with the common evaluation objective identified by each column. Table S17 presents the complete stepwise pathways for Entropy, CV, and CRITIC, including the selected scheme and cMIEI at every stage. Together, these additions make it possible to trace how the different weighting methods lead to their final configurations and to compare those configurations under explicitly defined evaluation objectives.

Finally, we have addressed the terminology and presentation issues raised in the minor comments. We have harmonized the variable symbols and wind-component definitions, clarified the cMIEI terminology and the distinct aggregation or display transformations used in the figures, revised expressions implying statistical significance where no corresponding test was reported, and strengthened the links between the parameterization discussion and the variable-level evidence. We have also clarified the conditional nature of the prescribed process order. These revisions are intended to make the methodological interpretation, supporting evidence, and scope of the conclusions clearer and more consistent across the manuscript and Supporting Information.

Our detailed point-by-point responses and the corresponding manuscript revisions 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.](#)

Major comment 1

1. The proposed multiplicative rule is not non-compensatory
P12, L248-275; Eqs. (13)-(15); related discussion at P29, L617-633.

The manuscript describes MRE x CV weighting as a "non-compensatory joint constraint." However, $w_1 = \text{softmax}(\text{MRE})$, $w_2 = \text{softmax}(\text{CV})$, and the normalized product is exactly $w = \text{softmax}(\text{MRE} + \text{CV})$. Thus, MRE and CV are additive before the final Softmax, and a high MRE can compensate for a low CV. The method should therefore not be described as strictly non-compensatory. Please revise the mathematical interpretation and related claims.

Figure 2 also suggests that the MRE component largely determines the final weights. Please add an MRE-only baseline, to the comparison in Fig. 8. This is needed to show whether the proposed method adds useful information beyond MRE alone.

Response:

We thank the reviewer for identifying these two related issues. We have corrected the interpretation of the combination rule and added the MRE-only comparison to determine how CV affects the evaluation. We address the mathematical interpretation first, followed by the evidence for the additional diagnostic information provided by CV.

(a) Mathematical interpretation and the additive alternative

We agree that the normalized product is compensatory and that the description “non-compensatory joint constraint” is incorrect. The rule combines simulation difficulty, characterized by MRE, with scheme-dependent variation in errors, characterized by CV. Using both diagnostics does not in itself preclude compensation between them. Its mathematical interpretation therefore requires distinguishing the diagnostic scores from the normalized component weights: the addition identified by the reviewer occurs before the final Softmax, whereas the additive alternative considered here averages two weight distributions after separate Softmax transformations. We therefore first establish the compensatory nature of the product rule through its equivalent expression, and then examine how the two combination rules differ in allocating weights.

Specifically, let $a_m = \overline{MRE}_m$ and $b_m = CV_{|ME|m}$, with $p_m = \text{softmax}(a)_m$ and $q_m = \text{softmax}(b)_m$. The normalized product is

$$\omega_m = \frac{p_m q_m}{\sum_j p_j q_j} = \frac{\exp(a_m + b_m)}{\sum_j \exp(a_j + b_j)}.$$

Thus, MRE and CV enter additively before the final Softmax, and an increase in one diagnostic score can offset a decrease in the other. In contrast, the equally weighted additive alternative combines the separately normalized component weights as $\omega_m^{add} = (p_m + q_m)/2$. The two rules are therefore both compensatory, but combine information at different stages: the product rule applies Softmax to the sum of diagnostic scores, whereas the additive rule averages the two component-weight distributions.

There is nevertheless a specific distinction in how the rules treat component support. For a fixed sum $s_m = p_m + q_m$,

$$p_m q_m = \frac{s_m^2 - (p_m - q_m)^2}{4}.$$

Consequently, among variables with the same additive weight, the product rule assigns a higher final weight to the variable whose two component weights are more balanced; the common normalization across variables preserves this ordering. This property provides a rationale for integrating the two diagnostics through a product. It does not require both original diagnos-

tic scores to be high. We have therefore replaced the claims of a non-compensatory or stronger joint constraint with this precise interpretation.

The practical evidence comes from the additive-rule sensitivity test (Table S10 and Sect. 3.5). The two rules selected the same schemes in all five Southeast China stages and in four of the five Northwest China stages, differing only at the Northwest cumulus stage. In the independent 2010 Northwest China validation, the multiplicative-rule configuration had a slightly lower cMIEI than the additive-rule configuration under the same reference weights (1.308 versus 1.312). This supports its use in the present comparison. The revised discussion distinguishes this case-specific evidence from the mathematical property above.

(b) Contribution of CV beyond the MRE-only baseline

The combination-rule comparison above addresses how the two diagnostics are integrated. The MRE-only baseline comparison addresses the complementary question of what changes when CV is included.

We have added an MRE-only baseline to Fig. 8 and performed a stage-by-stage comparison to quantify the contribution of CV. The baseline is $\omega_m^{\text{MRE}} = \text{softmax}(\overline{\text{MRE}})_m$. It uses the same candidate simulations, observations, variable grouping, sample handling, and cMIEI calculation as the proposed method, with only the variable weights changed.

The results confirm the reviewer's observation that MRE largely determines the final weights. The MRE-only average weights for PRE and UV10 were 0.239 and 0.445 in Northwest China and 0.307 and 0.357 in Southeast China; the corresponding proposed weights were 0.240 and 0.431, and 0.317 and 0.342. CV introduced stage-dependent adjustments, with a maximum absolute weight change of 0.048 for PRE in the Southeast cumulus stage.

We also recalculated the candidate-scheme rankings. Non-winning candidates exchanged adjacent ranks at positions 7 and 8 in the Southeast microphysics stage and positions 12 and 13 in the Southeast PBL/surface-layer stage. The stage-level weight changes and scheme selections are reported in Table S11 of the Supporting Information.

The additional diagnostic information provided by CV concerns the distinction between the mean error level of a variable and the relative differences in its errors across candidate schemes. MRE describes the former, whereas CV measures the relative dispersion of scheme-level mean absolute errors. Thus, two variables with the same mean relative error can still differ in their ability to distinguish candidate schemes within the physical-process group being evaluated. Including CV incorporates this distinction into variable weighting, allowing the evaluation to account for both simulation difficulty and scheme-dependent error response. The 0.048 increase in PRE weight in the Southeast cumulus stage and the changes in some candidate rankings demonstrate this component's actual influence on the present evaluation. Its diagnostic role does not require a change in the top-ranked scheme.

In the present cases, the relatively large errors in precipitation and wind give MRE a dominant role in weight allocation. As understanding and parameterization of the relevant physical processes improve, one scenario worth testing is whether differences in MRE across variables narrow while differences in CV associated with scheme-dependent error responses persist. Under these conditions, CV could have a more prominent relative influence on weight allocation. Where candidates have similar overall performance but different strengths across variables, such adjustments could also affect scheme selection. Future evaluations using more accurate simulations and different candidate sets could test this potential role.

Corresponding revisions:

Section 2.3.3 (page 13, lines 271–285)

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

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

Section 3.5 (page 25, lines 560–565; page 27, lines 595–597)

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

These results indicate that the main scheme-selection outcome was largely stable to the choice of combination rule, although the different weight allocations changed the cumulus-stage selection in Northwest China.

Section 3.5, Figure 8 caption (page 26)

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

Section 4.1 (page 30, lines 650–659; page 31, lines 660–686)

The resulting weights therefore jointly reflect simulation difficulty and the scheme-dependent response of variable errors within the evaluated physical-process group.

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

The additive-rule sensitivity test assesses the practical consequences of this combination choice. The additive test produced the same selected configuration as the multiplicative rule in Southeast China and changed only the cumulus-stage choice in Northwest China. This pattern indicates that the main screening outcome was largely stable across the two combination rules, while the Northwest China difference illustrates that their weight allocations can lead to different selections. In the independent 2010 Northwest China validation, the multiplicative-rule configuration had a slightly lower cMIEI than the additive-rule configuration (1.308 versus 1.312) under the same reference weights. These results support the use of the multiplicative combination in the present case studies.

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

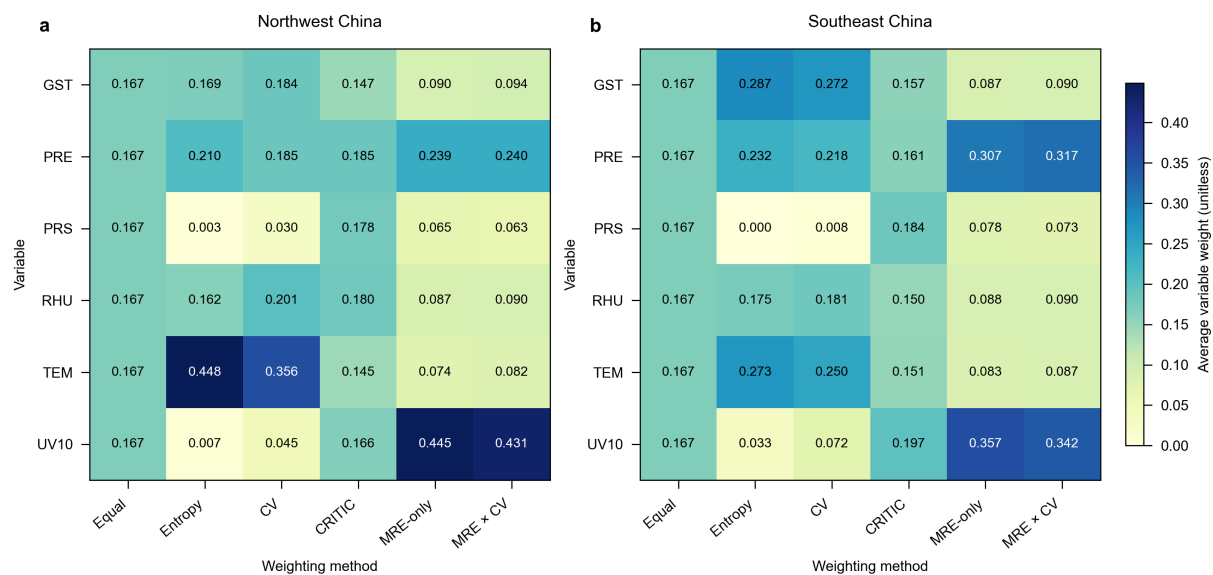


Figure 8. Comparison of average variable weights derived from equal weighting, entropy weighting, coefficient-of-variation weighting, CRITIC weighting, the MRE-only baseline, and the proposed MRE × CV weighting. Values represent averages across the five physical-process evaluation stages. The MRE-only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

Table S11 Absolute changes in variable weights and schemes selected under MRE-only and MRE
 × CV weighting

Region	Stage	Scheme selected under both rules	GST	PRE	PRS	RHU	TEM	UV10
Northwest	Microphysics	P3	0.0011	0.0073	0.0019	0.0009	0.0042	0.0116
	PBL/Surface Layer	YSU + MM5	0.0004	0.0039	0.0013	0.0066	0.0069	0.0087
	Cumulus	Tiedtke	0.0021	0.0158	0.0020	0.0004	0.0001	0.0122
	Radiation	GFDL+ GFDL	0.0081	0.0007	0.0024	0.0015	0.0099	0.0164
	LSM	Noah LSM	0.0146	0.0130	0.0041	0.0043	0.0200	0.0218
Southeast	Microphysics	WDM6	0.0007	0.0084	0.0024	0.0012	0.0012	0.0066
	PBL/Surface Layer	Shin-Hong + MM5	0.0042	0.0015	0.0032	0.0024	0.0030	0.0048
	Cumulus	Grell 3	0.0051	0.0480	0.0080	0.0024	0.0010	0.0315
	Radiation	GFDL	0.0089	0.0006	0.0066	0.0069	0.0172	0.0269
	LSM	Noah LSM	0.0098	0.0022	0.0046	0.0064	0.0004	0.0091

Major comment 2

2. A common variable-level comparison of all final configurations is needed

Figures 4 and 7 provide variable-level results only for the proposed and equal-weighting configurations. Entropy, CV, and CRITIC are compared only through integrated metrics. Lines 715-719 report RMSE reductions of 0.98 degrees C for TEM, 2.47 degrees C for GST, and 1.38 mm d-1 for PRE, but the reference configuration and absolute RMSE values are not given. It is also unclear why only these variables are reported.

Please compare the final configurations selected by all five methods using the same daily station samples and the same unweighted variable-level metrics, such as RMSE, MAE or mean bias, and correlation. This would show whether the proposed method improves its highly weighted target variables without unacceptable losses in the other variables.

Response:

We thank the reviewer for requesting a complete and directly comparable assessment of the selected configurations. We agree that a few RMSE reductions without explicit reference configurations and absolute values were insufficient to assess the balance between improvements and losses across variables. We have therefore compared the final configurations selected by Equal, $MRE \times CV$, Entropy, CV, and CRITIC using the same daily station samples and unweighted RMSE, MAE, mean bias, and Pearson correlation for GST, PRE, PRS, RHU, TEM, U10, and V10. The comparison covers June–August in Northwest 2007, Southeast 2009, and both regional validation cases in 2010. Supporting Information Table S15 provides the complete results, and Sections 2.3.1, 3.5, 3.6, and 4.2 now describe the comparison and its interpretation.

To make the individual differences explicit, the reference in the comparisons below is the configuration selected by equal weighting unless another method is named. The originally reported TEM and GST reductions refer to Northwest 2007: their RMSE values decreased from 3.686 to 2.704 °C and from 11.053 to 8.588 °C, respectively, giving reductions of 0.98 and 2.47 °C. These variables also improved in MAE, absolute mean bias, and correlation. Checking the precipitation comparison identified a statement requiring correction: in Southeast 2009, PRE RMSE was 18.027 mm d⁻¹ for Equal and 20.238 mm d⁻¹ for the proposed configuration, an increase of 2.21 mm d⁻¹ rather than the previously stated reduction of 1.38 mm d⁻¹. Section 3.5 now reports these absolute values within the full variable-level comparison. Section 4.2 interprets their meteorological relevance and the associated trade-offs.

For the high-weight target variables, the common comparison shows improvements in wind and, in some cases, precipitation. In Southeast 2009 and 2010, the proposed configuration reduced both RMSE and MAE for U10 and V10; V10 RMSE decreased from 2.261 to 2.143 m s⁻¹ and from 2.173 to 2.069 m s⁻¹, respectively. In Northwest 2010, precipitation RMSE and MAE decreased from 3.927 to 3.838 mm d⁻¹ and from 1.269 to 1.242 mm d⁻¹, respectively,

while both error metrics also decreased for U10 and V10. Its RMSE for precipitation and each wind component was also lower than that of the Entropy and CV configurations.

The effects on other variables differ in magnitude and across metrics. Across all four cases, the absolute pressure RMSE difference was less than 0.011 hPa, and the correlation difference was less than 0.001. In Northwest 2007, V10 RMSE increased from 2.616 to 2.677 m s⁻¹, while correlation increased from 0.211 to 0.248 and MAE decreased slightly from 1.991 to 1.988 m s⁻¹. Relative to the CV configuration in that case, precipitation RMSE increased from 3.571 to 3.657 mm d⁻¹, whereas MAE decreased from 1.277 to 1.176 mm d⁻¹. These examples distinguish small changes and trade-offs across error measures from a consistent deterioration across metrics.

The comparison also identifies larger costs: RHU in Northwest 2007, PRE and RHU in both Southeast cases, and GST and TEM in Northwest 2010 showed increases in both RMSE and MAE relative to Equal. For example, GST and TEM RMSE in Northwest 2010 increased from 8.403 and 3.352 to 10.779 and 4.129 °C, respectively. Thus, improvements in several high-weight targets coexist with small changes, cross-metric trade-offs, and some larger error increases in other variables. Table S15 makes the magnitudes of these losses explicit alongside the gains for all five methods, providing the evidence needed to assess the balance for the intended application. The revised Results and Discussion interpret the selected configurations in terms of this multivariable balance under the prescribed integrated criterion, and the meteorological interpretation is linked to the observed temperature and wind results.

Corresponding revisions:

Section 2.3.1 (page 10, lines 202–204)

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

Section 3.5 (page 27, lines 598–600; page 28, lines 601–604)

Variable-level performance differed between the two screening regions (Table S15). In Northwest 2007, the proposed configuration improved RMSE, MAE, absolute mean bias, and correlation for GST and TEM relative to Equal. TEM RMSE decreased from 3.686 to 2.704 °C, and GST RMSE decreased from 11.053 to 8.588 °C. The respective reductions were 0.98 and 2.47 °C, whereas RHU RMSE and MAE increased. Relative to the CV configuration in the same case, precipitation RMSE was slightly higher but MAE was lower.

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

Section 3.6 (page 29, lines 616–619)

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

Section 4.2 (page 34, lines 763–770)

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

Table S15 Unweighted daily station-level performance of the five final configurations**(a) Northwest China, 2007**

Variable	Unit	Method	RMSE	MAE	Mean Bias	r
GST	°C	Equal	11.053	9.893	-9.740	0.647
		MRE × CV	8.588	7.386	-6.791	0.653
		Entropy	7.791	6.713	-6.002	0.693
		CV	9.281	8.221	-7.930	0.705
		CRITIC	11.053	9.893	-9.740	0.647
PRE	mm d ⁻¹	Equal	3.522	1.165	0.021	0.438
		MRE × CV	3.657	1.176	0.002	0.355
		Entropy	3.660	1.258	0.107	0.381
		CV	3.571	1.277	0.178	0.436
		CRITIC	3.522	1.165	0.021	0.438
PRS	hPa	Equal	13.298	7.078	-5.719	0.983
		MRE × CV	13.303	7.100	-5.731	0.983
		Entropy	13.502	7.451	-6.452	0.983
		CV	13.533	7.503	-6.539	0.983
		CRITIC	13.298	7.078	-5.719	0.983
RHU	%	Equal	16.011	12.899	-10.379	0.717
		MRE × CV	19.165	15.859	-14.555	0.696
		Entropy	16.697	13.619	-11.509	0.749
		CV	15.063	12.072	-9.013	0.760
		CRITIC	16.011	12.899	-10.379	0.717
TEM	°C	Equal	3.686	3.040	-2.529	0.858
		MRE × CV	2.704	2.173	-0.495	0.874
		Entropy	2.587	2.071	-0.427	0.886
		CV	2.915	2.370	-1.527	0.886
		CRITIC	3.686	3.040	-2.529	0.858
U10	m s ⁻¹	Equal	2.784	2.166	-0.257	0.371
		MRE × CV	2.820	2.214	-0.577	0.459
		Entropy	3.117	2.399	-0.258	0.456
		CV	3.115	2.393	-0.181	0.433
		CRITIC	2.784	2.166	-0.257	0.371
V10	m s ⁻¹	Equal	2.616	1.991	-0.222	0.211
		MRE × CV	2.677	1.988	-0.236	0.248
		Entropy	2.980	2.215	-0.540	0.247
		CV	3.086	2.299	-0.471	0.238
		CRITIC	2.616	1.991	-0.222	0.211

(b) Southeast China, 2009

Variable	Unit	Method	RMSE	MAE	Mean Bias	r
GST	°C	Equal	3.526	2.843	-0.156	0.607
		MRE × CV	3.627	2.894	-0.755	0.600
		Entropy	3.551	2.865	-0.156	0.616
		CV	4.273	3.437	-2.154	0.578
		CRITIC	3.526	2.843	-0.156	0.607
PRE	mm d ⁻¹	Equal	18.027	8.358	1.450	0.486
		MRE × CV	20.238	11.482	6.136	0.469
		Entropy	18.905	8.091	0.735	0.449
		CV	18.047	9.983	4.219	0.496
		CRITIC	18.027	8.358	1.450	0.486
PRS	hPa	Equal	20.577	10.958	-0.868	0.789
		MRE × CV	20.581	10.948	-0.851	0.789
		Entropy	20.588	11.006	-0.956	0.789
		CV	20.569	10.935	-0.863	0.790
		CRITIC	20.577	10.958	-0.868	0.789
RHU	%	Equal	9.693	7.740	3.084	0.525
		MRE × CV	11.206	9.068	5.632	0.487
		Entropy	9.225	7.311	-0.596	0.551
		CV	11.650	9.433	5.467	0.398
		CRITIC	9.693	7.740	3.084	0.525
TEM	°C	Equal	1.882	1.420	-0.277	0.744
		MRE × CV	2.020	1.589	-0.675	0.732
		Entropy	1.846	1.370	0.043	0.753
		CV	2.360	1.922	-1.221	0.703
		CRITIC	1.882	1.420	-0.277	0.744
U10	m s ⁻¹	Equal	1.701	1.264	0.164	0.505
		MRE × CV	1.670	1.223	0.039	0.491
		Entropy	1.719	1.268	0.209	0.493
		CV	1.658	1.201	0.004	0.480
		CRITIC	1.701	1.264	0.164	0.505
V10	m s ⁻¹	Equal	2.261	1.741	0.547	0.558
		MRE × CV	2.143	1.636	0.430	0.566
		Entropy	2.241	1.736	0.587	0.562
		CV	2.104	1.604	0.202	0.552
		CRITIC	2.261	1.741	0.547	0.558

(c) Northwest China, 2010

Variable	Unit	Method	RMSE	MAE	Mean Bias	r
GST	°C	Equal	8.403	7.331	-6.501	0.630
		MRE × CV	10.779	9.722	-9.442	0.636
		Entropy	7.628	6.610	-5.760	0.678
		CV	9.139	8.141	-7.727	0.685
		CRITIC	8.403	7.331	-6.501	0.630
PRE	mm d ⁻¹	Equal	3.927	1.269	0.084	0.063
		MRE × CV	3.838	1.242	0.037	0.064
		Entropy	3.879	1.306	0.156	0.078
		CV	3.938	1.324	0.166	0.083
		CRITIC	3.927	1.269	0.084	0.063
PRS	hPa	Equal	13.588	7.680	-5.926	0.982
		MRE × CV	13.578	7.660	-5.897	0.982
		Entropy	13.820	7.963	-6.637	0.983
		CV	13.844	8.008	-6.734	0.983
		CRITIC	13.588	7.680	-5.926	0.982
RHU	%	Equal	20.635	16.685	-13.821	0.518
		MRE × CV	18.060	14.338	-9.672	0.541
		Entropy	18.251	14.838	-10.924	0.608
		CV	17.276	13.895	-8.568	0.611
		CRITIC	20.635	16.685	-13.821	0.518
TEM	°C	Equal	3.352	2.674	-0.461	0.804
		MRE × CV	4.129	3.458	-2.483	0.787
		Entropy	3.178	2.545	-0.414	0.825
		CV	3.478	2.855	-1.523	0.819
		CRITIC	3.352	2.674	-0.461	0.804
U10	m s ⁻¹	Equal	3.292	2.563	-0.429	0.297
		MRE × CV	3.070	2.397	-0.058	0.277
		Entropy	3.589	2.730	-0.042	0.331
		CV	3.517	2.685	0.004	0.320
		CRITIC	3.292	2.563	-0.429	0.297
V10	m s ⁻¹	Equal	2.952	2.176	-0.457	0.112
		MRE × CV	2.825	2.123	-0.409	0.118
		Entropy	3.188	2.358	-0.713	0.155
		CV	3.210	2.387	-0.672	0.160
		CRITIC	2.952	2.176	-0.457	0.112

(d) Southeast China, 2010

Variable	Unit	Method	RMSE	MAE	Mean Bias	r
GST	°C	Equal	3.699	2.948	-0.732	0.696
		MRE × CV	3.875	3.043	-1.225	0.685
		Entropy	3.621	2.898	-0.715	0.713
		CV	4.655	3.702	-2.636	0.654
		CRITIC	3.699	2.948	-0.732	0.696
PRE	mm d ⁻¹	Equal	19.264	8.958	1.125	0.454
		MRE × CV	20.671	11.484	5.065	0.452
		Entropy	19.312	8.557	0.129	0.451
		CV	19.368	10.732	3.725	0.430
		CRITIC	19.264	8.958	1.125	0.454
PRS	hPa	Equal	20.691	10.994	-0.862	0.789
		MRE × CV	20.693	10.972	-0.821	0.789
		Entropy	20.700	11.042	-0.950	0.789
		CV	20.678	10.964	-0.854	0.790
		CRITIC	20.691	10.994	-0.862	0.789
RHU	%	Equal	9.714	7.871	3.420	0.539
		MRE × CV	11.299	8.996	6.007	0.497
		Entropy	8.931	7.254	0.044	0.577
		CV	11.681	9.394	5.697	0.419
		CRITIC	9.714	7.871	3.420	0.539
TEM	°C	Equal	1.976	1.541	-0.413	0.805
		MRE × CV	2.124	1.673	-0.763	0.793
		Entropy	1.905	1.467	-0.132	0.813
		CV	2.487	2.036	-1.324	0.767
		CRITIC	1.976	1.541	-0.413	0.805
U10	m s ⁻¹	Equal	1.747	1.293	0.029	0.517
		MRE × CV	1.696	1.246	-0.107	0.527
		Entropy	1.787	1.330	0.125	0.504
		CV	1.697	1.250	-0.138	0.505
		CRITIC	1.747	1.293	0.029	0.517
V10	m s ⁻¹	Equal	2.173	1.706	0.548	0.571
		MRE × CV	2.069	1.619	0.425	0.575
		Entropy	2.163	1.703	0.603	0.572
		CV	2.060	1.601	0.195	0.559
		CRITIC	2.173	1.706	0.548	0.571

Minor comment 1

L175 and subsequent figures and text.

The seven variables are introduced as T2, GST, PRS, PRE, RH, U10, and V10, but later figures use GST, PRE, PRS, RHU, TEM, and UV10. Please use consistent names and define all regrouping.

Response:

We thank the reviewer for pointing out the inconsistent notation. We have standardized the manuscript to use TEM for 2 m air temperature and RHU for relative humidity, consistently with the figures and tables. We now define the seven evaluated variables at first use and explicitly distinguish them from the six weighting groups. GST, PRE, PRS, RHU, and TEM form five individual groups, while U10 and V10 form the joint wind-vector group UV10 and share one group weight. U10 and V10 are retained separately in component-level scatter plots and tables. References to these components and their joint group now use wind-component or wind-vector terminology. The corresponding definitions have been added to Section 2.2.

Corresponding revisions:

Section 2.2 (page 9, lines 174–180)

The seven evaluated variables were 2 m air temperature (TEM), ground surface temperature (GST), air pressure (PRS), precipitation (PRE), relative humidity (RHU), and the zonal (U10) and meridional (V10) components of 10 m wind.

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

Minor comment 2

L175-176. The manuscript defines U10 as meridional wind and V10 as zonal wind. Under the usual WRF convention, U is zonal and V is meridional. Please check and correct these definitions.

Response:

We thank the reviewer for pointing this out. We have corrected the definitions in Section 2.2: U10 denotes the zonal component of 10 m wind, and V10 denotes the meridional component.

Corresponding revisions:**Section 2.2 (page 9, lines 174–176)**

The seven evaluated variables were 2 m air temperature (TEM), ground surface temperature (GST), air pressure (PRS), precipitation (PRE), relative humidity (RHU), and the zonal (U10) and meridional (V10) components of 10 m wind.

Minor comment 3

Around L287 and throughout the manuscript. Please clearly define MIEI, cMIEI, and the median-normalized MIEI shown in Figs. 3 and 6. Their notation and roles should remain consistent throughout the text, equations, tables, and captions.

Response:

We thank the reviewer for highlighting the need to distinguish these quantities. We have revised Section 2.3.2 to define MIEI as the name of the evaluation index and cMIEI as the centered form used in this study. Centering means subtracting the respective means from the simulated and observed series before calculating their amplitude and pattern statistics. Equation (5) now explicitly uses cMIEI, and the definitions of the standard-deviation ratio and integrated cVSC are given alongside it. We use cMIEI consistently for the calculated index in the text and table headings.

For the heatmaps in Figs. 3(a) and 6(a), median normalization expresses cMIEI as a relative deviation from its reference median:

$$\widetilde{\text{cMIEI}}_n = \frac{\text{cMIEI}_n - C_{\text{med}}}{C_{\text{med}}}, \quad C_{\text{med}} = \text{median}_{j \in \mathcal{E}}(\text{cMIEI}_j).$$

Here, the reference ensemble $\mathcal{E}$ pools configurations across the five physical-process stages within each region, retaining only the first occurrence of a repeated stage-entry baseline. The median is calculated separately for each variable column and the integrated column, so each column is evaluated relative to its own reference median. Zero denotes the reference median, negative values indicate lower cMIEI, and positive values indicate higher cMIEI.

This transformation displays the relative performance of the individual variables and the integrated index on a common color scale. Because the reference median is positive, the trans-

formation preserves the numerical ordering within each series. Scheme selection and numerical comparisons use cMIEI before this display transformation. We have briefly explained the transformation and its purpose in the corresponding captions.

The existing MIEI headings in Supporting Information Tables S2 and S3 have been changed to cMIEI. These revisions clarify the metric definitions and standardize the notation without altering the calculation method or the scheme-selection results.

Corresponding revisions:

Section 2.3.2 (page 10, lines 212–218; page 11, lines 221–225)

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

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

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

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

Section 3.1, Figure 3 caption (page 17)

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

Section 3.3, Figure 6 caption (page 22)

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

Supporting Information, Tables S2 and S3 (pages 3–4)

The existing index column heading is now “cMIEI”.

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	cMIEI
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	cMIEI
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

Minor comment 4

L425. The word "significantly" requires a statistical significance test. Please provide such a test or replace it with a descriptive term such as "substantially" or "noticeably."

Response:

We thank the reviewer for pointing this out. We have replaced "significantly higher" with "higher" in the description of Fig. 4. The revised sentence describes the observed differences without implying statistical significance.

We have also revised related wording: "significantly affects" is now "affects" in the CLM discussion, "significant errors" is now "large errors," and "significant fluctuations and variability" is now "substantial variation." In the description of the 2010 extreme-year conditions, "significant flooding" has been replaced with "severe flooding" to describe the severity of the event more precisely. The reported numerical results are unchanged.

Corresponding revisions:

Section 3.2 (page 19, lines 437–438; page 20, line 439)

However, the R^2 values of the configuration selected by the proposed weighting scheme reached 0.41 and 0.15, which were higher than those of the equal weight scheme (0.20 and 0.12), indicating better evaluated agreement for 10 m wind.

Section 3.1 (page 15, lines 367–368; page 16, line 369)

Although CLM, as a land surface model, does not directly govern precipitation-related processes, its influence on the hydrological cycle and land-atmosphere interactions affects precipitation outcomes.

Section 3.1 (page 15, lines 351–353)

The higher weights for 10 m wind and precipitation indicate that all physical processes in the WRF model have large errors in simulating these variables, reflecting the model's weaker simulation capability for these variables.

Section 3.3 (page 21, lines 465–466)

Surface temperature, air temperature, and relative humidity exhibited substantial variation in cMIEI values throughout the evaluation.

Section 2.2 (page 6, lines 136–137)

Concurrently, southern China experienced heavy rainfall in June and August, leading to severe flooding in the coastal region of Southeast, with substantial loss of life and property.

Minor comment 5

Figs. 2, 5, and 8 and related text. The figures combine U10 and V10 as UV10, while the text often calls both components "wind speed." Please clarify whether they enter MIEI as a two-dimensional vector, are combined into wind-speed magnitude, or are treated as separate zonal and meridional components.

Response:

We thank the reviewer for highlighting this ambiguity. UV10 enters the evaluation as a two-dimensional 10 m wind vector composed of the zonal component U10 and the meridional component V10. It is not reduced to the scalar wind-speed magnitude, and the two components are not assigned independent weights in the integrated evaluation. Accordingly, Figs. 2, 5, and 8 show one UV10 weight among the six evaluation groups.

This treatment follows the vector-field evaluation framework already cited in the manuscript. Xu et al. (2016) evaluate vector fields jointly using the vector similarity coefficient (VSC), root mean square length (RMSL), and root mean square vector difference (RMSVD), and describe their centered counterparts for vector anomaly fields. Zhang et al. (2021) extend this framework to integrated evaluation of scalar and vector variables: a scalar variable has one component, whereas a two-dimensional wind variable has two components that are normalized together. Their Fig. 3 illustrates this distinction using paired wind components, and their MVIETool identifies a vector variable by enclosing its component names in parentheses. Our evaluation uses this grouping for (u10, v10).

Specifically, let $\mathbf{f}_t = (u_{f,t}, v_{f,t})$ and $\mathbf{r}_t = (u_{r,t}, v_{r,t})$ denote the simulated and observed wind vectors at a valid daily station sample. In the centered evaluation, $\mathbf{f}'_t = \mathbf{f}_t - \bar{\mathbf{f}}$ and $\mathbf{r}'_t = \mathbf{r}_t - \bar{\mathbf{r}}$, where each component has its own sample mean removed. The wind statistics are

$$L_{f,c} = \sqrt{\frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t\|^2}, \quad L_{r,c} = \sqrt{\frac{1}{T} \sum_{t=1}^T \|\mathbf{r}_t\|^2},$$

$$\text{cVSC}_{\text{UV10}} = \frac{\frac{1}{T} \sum_{t=1}^T \mathbf{f}_t \cdot \mathbf{r}_t}{L_{f,c} L_{r,c}}, \quad \text{cRMSVD}_{\text{UV10}} = \sqrt{\frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t - \mathbf{r}_t\|^2}.$$

Thus, the inner product contains both $u'_{f,t}u'_{r,t}$ and $v'_{f,t}v'_{r,t}$, and the squared vector difference contains the squared errors of both centered components. These operations retain the joint component information, including differences in the direction and amplitude of the centered wind vectors; computing statistics from $\sqrt{u^2 + v^2}$ alone would discard directional information. The UV10 amplitude ratio in cMIEI is $L_{f,c}/L_{r,c}$. When constructing the integrated field, both wind components use the same observed RMSL scale and the same group weight, preserving their

treatment as one vector variable. The observed RMSL used for this scaling is calculated from the original wind vectors, whereas the centered statistics use component-wise anomalies.

The separate U10 and V10 scatter plots and the component-level results in Supporting Information Table S15 provide complementary diagnostics of the two components. They do not change the vector grouping used for weighting and integrated evaluation. We have clarified this distinction in Sections 2.2 and 2.3 and used “10 m wind,” “zonal component,” and “meridional component” where appropriate. These clarifications describe the existing implementation without changing the calculations or scheme-selection results.

Corresponding revisions:

Section 2.2 (page 9, lines 177–180)

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

Section 2.3.1 (page 10, lines 200–201)

For UV10, the products and squares in Eqs. (1)–(4) are interpreted as vector inner products and squared Euclidean norms, respectively, with each wind component centered separately.

Section 2.3.2 (page 11, lines 221–225)

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

Minor comment 6

L672-705. The discussion mainly explains why the selected schemes are physically plausible based on their design. It does not show why these mechanisms made them outperform the competing schemes in this study. Please link the interpretation to direct variable-level results where possible, or present these explanations more cautiously as hypotheses.

Response:

We thank the reviewer for this comment. We have expanded Section 4.2 to explain how the physical characteristics of the selected schemes could contribute to their variable-level performance relative to competing configurations at the same screening stage. The explanations draw on the original parameterization studies and are presented as physical hypotheses. The specific points are as follows.

For Northwest microphysics, P3 predicts evolving ice properties, including rimed ice mass and volume. These properties affect particle density, fall speed, and sedimentation, allowing the growth and fallout of mixed-phase hydrometeors to respond to cloud conditions (Morrison and Milbrandt, 2015). This can alter how much precipitation reaches the surface and its distribution. In a region with intermittent precipitation and complex terrain, this flexibility is a possible contributor to P3's lower precipitation cMIEI than Thompson (1.090 versus 1.192).

For Southeast microphysics, WDM6 predicts cloud-water and rain number concentrations in addition to their mass, together with cloud condensation nuclei. Changes in rain number concentration affect particle size distributions and evaporation, and hence the fraction of precipitation reaching the surface (Lim and Hong, 2010). In the warm, moist Southeast environment, a more suitable balance among these processes could explain its lower precipitation cMIEI than P3 (1.099 versus 1.133). The two regional results thus suggest different possible microphysical pathways to the observed precipitation performance.

For the PBL/surface-layer combinations, YSU's non-local mixing and explicit PBL-top entrainment redistribute heat and momentum through the mixed layer and regulate exchange with the air above it (Hong et al., 2006). Changes in these exchanges affect near-surface temperature and wind profiles. This provides a physical explanation for why the Northwest YSU-MM5 combination could improve both UV10 and TEM relative to MYJ-MYJ, as indicated by their cMIEI values of 1.203 versus 1.207 and 0.528 versus 0.533. In Southeast China, Shin-Hong's local and non-local heat-transport terms alter the vertical thermal structure and surface-atmosphere temperature gradient (Shin and Hong, 2015). These effects on surface-atmosphere exchange may contribute to the lower TEM and GST cMIEI of Shin-Hong-MM5 relative to MYJ-MYJ (0.793 versus 0.805 and 0.960 versus 0.977), particularly in a coastal environment with land-sea thermal contrasts. Precipitation cMIEI changes little (1.099 to 1.100), so the clearest improvement in this comparison concerns thermal conditions.

For Southeast cumulus parameterization, closure assumptions determine how convective mass flux responds to the resolved atmospheric state, and the resulting heating and moistening tendencies affect precipitation development. Grell 3 uses an ensemble-based formulation built on the approaches of Grell (1993) and Grell and Devenyi (2002), sampling different closure and convective responses. This flexibility may better accommodate the range of convective conditions in the Southeast case, offering a possible explanation for its lower precipitation cMIEI than New Tiedtke (1.053 versus 1.068). Because precipitation has a high weight at this stage, this variable-level advantage contributes directly to its selection.

For radiation and land-surface processes, differences in radiative heating alter atmospheric stability and near-surface thermal gradients, while soil moisture and surface fluxes affect the partitioning of energy into sensible and latent heat. Their coupling can influence boundary-layer development, wind, and precipitation. This provides a physical pathway through which GFDL could contribute to the lower UV10 cMIEI in Northwest China and lower PRE and UV10 cMIEI in Southeast China. Noah's soil and surface-exchange processes and Tiedtke's sub-grid convective transport provide complementary roles within the selected combinations. These process-based explanations describe possible reasons for the observed performance without assigning the gains to an individually diagnosed mechanism.

Corresponding revisions:

The specific manuscript revisions are provided below.

Section 4.2 (page 32, lines 695–726; page 33, lines 727–760; page 34, lines 761–762)

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

In the arid region of Northwest China, limited water-vapor supply, complex terrain, strong land-atmosphere coupling, and large diurnal temperature variations make precipitation, near-surface wind, and temperature relevant to the evaluation of cloud microphysics, boundary-layer mixing, radiation, and surface energy-water exchanges. At the microphysics stage, P3 has the lowest precipitation cMIEI among the tested candidates, with a value of 1.090 compared with 1.192 for Thompson (Fig. 3a). P3 predicts bulk ice particle properties, including rimed ice mass and volume, allowing ice properties to evolve more continuously than in schemes with fixed ice categories (Morrison and Milbrandt, 2015). These evolving properties affect particle fall speeds

and sedimentation, providing a flexible treatment of mixed-phase hydrometeors. Changes in the growth and fallout of ice particles can alter the amount and distribution of precipitation reaching the surface, offering a possible explanation for P3's lower precipitation cMIEI in this region, where intermittent precipitation is sensitive to cloud microphysical processes.

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

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

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

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

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

The selected configurations thus exhibit variable-level performance that is physically interpretable under the contrasting regional climate backgrounds. The mechanisms discussed above remain hypotheses for these responses, rather than direct process attribution. Process-level diagnostics and multi-year validation would help establish how these mechanisms operate and whether the same combinations remain favorable across different circulation patterns and types of extreme events.

Minor comment 7

L780-795. This section repeats the limitations of the stepwise search and gives a long list of fixed settings and untested uncertainty sources. Please shorten it to state that the study isolates parameterization selection under a fixed WRF configuration and does not provide a complete uncertainty decomposition.

Response:

We thank the reviewer for this suggestion. We have shortened the end of Section 4.3 by condensing the repeated statement about order dependence to one sentence and removing the lengthy lists of fixed settings, untested uncertainty sources, and additional sensitivity experiments. The revised text states that the experiments examine physical parameterization choices with the remaining WRF settings held fixed, and that the $\text{MRE} \times \text{CV}$ framework is a reproducible screening tool rather than a complete uncertainty decomposition.

Corresponding revisions:

Section 4.3 (page 35, lines 820–825)

Nevertheless, the selected configuration remains conditional on the prescribed process order, because choices made at earlier stages determine the background configuration used to evaluate candidates at subsequent stages.

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

Minor comment 8

Tables 3 and 4 would benefit from further clarification. If each cMIEI value was calculated using its corresponding method-specific weights, the values reflect different evaluation objectives and may be difficult to compare directly. To improve comparability, the authors could evaluate all selected final configurations using the same reference weights or a common set of unweighted metrics. Alternatively, a cross-evaluation matrix could be provided, in which every stepwise-selected configuration is evaluated under each weighting method and the corresponding cMIEI values are reported.

Response:

We thank the reviewer for this suggestion. We agree that the final configurations should be compared under a shared evaluation objective. In Tables 3 and 4, all configurations within a given region-year were already evaluated using the same $\text{MRE} \times \text{CV}$ reference-weight vector. The method names identify how the configurations were selected, rather than a different scoring weight for each row. We have clarified the common reference weights in Sections 3.5 and 3.6. For Tables 3 and 4, the reference-weight pool comprised the five method-labelled final configurations, with the additive-rule configuration additionally included for Northwest 2010.

We have also added a complete cross-evaluation matrix. For each region-year, we calculated Equal, $\text{MRE} \times \text{CV}$, Entropy, CV, and CRITIC evaluation weights from the same five method-labelled final configurations and applied each vector to all five configurations. The 2010 evaluation weights were recalculated from the corresponding 2010 data. Thus, each column provides a comparison under one common evaluation objective. Table S16 reports all 100 cMIEI values.

The proposed configuration had the lowest cMIEI under the common $\text{MRE} \times \text{CV}$ objective in all four region-year cases, including 1.3082 in Northwest 2010 and 1.0052 in Southeast 2010. Favorable results extended to other evaluation objectives: the proposed configuration ranked first under CV weights in Northwest 2007, CRITIC weights in Southeast 2009, and both Equal and CRITIC weights in Southeast 2010. Alternative priorities favored other configurations. In Northwest 2010, Entropy ranked first under Equal, Entropy, and CV weights, while the CV-selected configuration ranked first under CRITIC weights. In Southeast 2009, the CV-selected configuration ranked first under Equal, Entropy, and CV weights.

These results clarify the practical role of the proposed method: it selects a configuration suited to an objective that jointly emphasizes simulation difficulty and scheme-dependent error response. The cross-evaluation shows how the preferred configuration changes with variable priorities. Together with the common unweighted RMSE, MAE, mean bias, and correlation results in Table S15, this provides both an integrated comparison under explicit objectives and a direct assessment of the meteorological improvements and trade-offs requested in Major comment 2.

The manuscript has been revised as follows.

Corresponding revisions:**Section 3.5 (page 26, lines 581–582)**

The final configurations in Tables 3 and 4 were evaluated using a common $\text{MRE} \times \text{CV}$ reference-weight vector within each region-year, calculated from the corresponding final-configuration error matrix.

Section 3.5, Table 3 caption (page 27)

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

Section 3.6 (page 28, lines 606–609)

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

Section 3.6, Table 4 caption (page 29)

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

Section 3.6 (page 29, lines 620–625)

Cross-evaluation using weights recalculated for each region-year retained the proposed configuration as the lowest-cMIEI configuration under $MRE \times CV$ weighting in both 2010 regions (1.3082 in Northwest China and 1.0052 in Southeast China; Table S16). In Southeast 2010, it also ranked first under Equal and CRITIC evaluation weights. Other objectives favored alternative configurations, including Entropy in Northwest 2010 under Equal, Entropy, and CV weights, and the CV-selected configuration under CRITIC weights. These differences reflect the variable priorities encoded in each evaluation objective and complement the unweighted variable-level comparisons in Table S15.

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

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

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

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

Table S16 Cross-evaluation of the five final configurations under common weighting objectives**(a) Northwest China, 2007**

Configuration	Equal	MRE $\times$ CV	Entropy	CV	CRITIC
Equal	1.1242	1.1863	0.8099	1.0191	1.1060
MRE $\times$ CV	1.1300	1.1663	0.7914	1.0064	1.1109
Entropy	1.1367	1.1982	0.7841	1.0129	1.1162
CV	1.1306	1.2069	0.7918	1.0188	1.1108
CRITIC	1.1242	1.1863	0.8099	1.0191	1.1060

(b) Southeast China, 2009

Configuration	Equal	MRE $\times$ CV	Entropy	CV	CRITIC
Equal	1.0148	1.0094	1.0252	1.0245	1.0032
MRE $\times$ CV	1.0160	1.0061	1.0378	1.0361	1.0007
Entropy	1.0236	1.0175	1.0538	1.0505	1.0054
CV	1.0133	1.0079	1.0131	1.0130	1.0065
CRITIC	1.0148	1.0094	1.0252	1.0245	1.0032

(c) Northwest China, 2010

Configuration	Equal	MRE $\times$ CV	Entropy	CV	CRITIC
Equal	1.2821	1.3123	0.9361	1.1342	1.2452
MRE $\times$ CV	1.2815	1.3082	0.9347	1.1309	1.2405
Entropy	1.2725	1.3249	0.9116	1.1202	1.2294
CV	1.2727	1.3220	0.9163	1.1236	1.2293
CRITIC	1.2821	1.3123	0.9361	1.1342	1.2452

(d) Southeast China, 2010

Configuration	Equal	MRE $\times$ CV	Entropy	CV	CRITIC
Equal	1.0254	1.0142	1.0569	1.0537	1.0074
MRE $\times$ CV	1.0204	1.0052	1.0587	1.0549	0.9994
Entropy	1.0213	1.0137	1.0504	1.0468	1.0056
CV	1.0414	1.0298	1.0795	1.0754	1.0204
CRITIC	1.0254	1.0142	1.0569	1.0537	1.0074

Note. Rows identify the configuration-selection method; columns identify the evaluation weighting rule. For each region-year, all five weight vectors were derived from the same five method-labelled final configurations. Methods selecting the same configuration retain their separate labels. Each column applies one common vector to all configurations using common daily June–August station samples. Values are dimensionless cMIEI; bold marks the column minimum. Compare configurations within each column.

Minor comment 9

It would also be helpful to present the complete stepwise pathways for entropy, CV, and CRITIC weighting. Reporting the selected scheme and cMIEI at each stage would make the effects of different weighting methods on the selection pathways and final configurations easier to compare.

Response:

We thank the reviewer for this suggestion. We have added Supporting Information Table S17 to present the complete stepwise selection pathways for Entropy, CV, and CRITIC weighting in Northwest China (2007) and Southeast China (2009). The table reports the selected scheme and its cMIEI at each of the five stages, covering all 30 method–region–stage selections. The stages follow the prescribed order of microphysics, PBL/surface layer, cumulus, radiation, and land surface. Each stage’s selected configuration is carried forward to the next stage, so the five selections specify the final configuration for each method.

The side-by-side presentation makes the pathway differences explicit. In Northwest China, Entropy and CV both select P3, YSU + MM5, New Tiedtke, and NoahMP, but select GFDL and RRTMG radiation, respectively. CRITIC instead retains MYJ + MYJ, Tiedtke, and Noah LSM. In Southeast China, Entropy selects ACM2 + MM5, whereas CV and CRITIC select Shin-Hong + MM5. CV selects NSAS convection and New Goddard radiation, while Entropy and CRITIC select New Tiedtke and GFDL. All three methods retain WDM6 microphysics and Noah LSM in Southeast China. These differences show which individual choices distinguish the final configurations.

The cMIEI values in Table S17 are the original stagewise selection scores calculated using the weights of the corresponding method and stage. The pathways complement the common-reference final-configuration comparison in Table 3, the unweighted variable-level results in Table S15, and the cross-evaluation matrix in Table S16. A brief reference to Table S17 has also been added to Sect. 3.5.

Corresponding revisions:

Section 3.5 (page 25, lines 539–541)

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

Table S17 Complete stepwise selection pathways for Entropy, CV and CRITIC weighting**(a) Northwest China, 2007**

Stage	Entropy Selected scheme	cMIEI	CV Selected scheme	cMIEI	CRITIC Selected scheme	cMIEI
Microphysics	P3	1.0143	P3	1.0276	P3	1.1501
PBL/surface layer	YSU + MM5	0.6666	YSU + MM5	0.7670	MYJ + MYJ	1.1264
Cumulus	New Tiedtke	1.1010	New Tiedtke	1.0887	Tiedtke	1.0986
Radiation	GFDL + GFDL	0.6574	RRTMG + RRTMG	0.8729	RRTMG + RRTMG	1.0986
Land surface	NoahMP	0.5777	NoahMP	0.7857	Noah LSM	1.0863

(b) Southeast China, 2009

Stage	Entropy Selected scheme	cMIEI	CV Selected scheme	cMIEI	CRITIC Selected scheme	cMIEI
Microphysics	WDM6	1.0968	WDM6	1.0909	WDM6	1.0504
PBL/surface layer	ACM2 + MM5	1.0106	Shin-Hong + MM5	1.0444	Shin-Hong + MM5	1.0492
Cumulus	New Tiedtke	1.0914	NSAS	1.0527	New Tiedtke	1.0054
Radiation	GFDL	0.9349	New Goddard	0.9981	GFDL	1.0165
Land surface	Noah LSM	0.9618	Noah LSM	1.0093	Noah LSM	1.0156

References cited in the responses

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.

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.

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.

Hong, S.-Y., Noh, Y., and Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes, *Mon. Weather Rev.*, 134, 2318-2341, doi:10.1175/MWR3199.1, 2006.

Jia, W., Zhang, X., Wang, H., Wang, Y., Wang, D., Zhong, J., Zhang, W., Zhang, L., Guo, L., and Lei, Y.: Comprehensive evaluation of typical planetary boundary layer (PBL) parameterization schemes in China—Part 1: Understanding expressiveness of schemes for different regions from the mechanism perspective, *Geosci. Model Dev.*, 16, 6635-6670, doi:10.5194/gmd-16-6635-2023, 2023.

Lim, K.-S. S. and Hong, S.-Y.: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models, *Mon. Weather Rev.*, 138, 1587-1612, doi:10.1175/2009MWR2968.1, 2010.

Liu, D., Yang, B., Zhang, Y., Qian, Y., Huang, A., Zhou, Y., and Zhang, L.: Combined impacts of convection and microphysics parameterizations on the simulations of precipitation and cloud properties over Asia, *Atmos. Res.*, 212, 172-185, doi:10.1016/j.atmosres.2018.05.017, 2018.

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.

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.

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.

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:ACMF SF>2.0.CO;2, 1989.

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.

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.

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.

Zhang, C., Wang, Y., and Hamilton, K.: Improved representation of boundary layer clouds over the Southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme, *Mon. Weather Rev.*, 139, 3489-3513, doi:10.1175/MWR-D-10-05091.1, 2011.

Zhang, M., Xu, Z., Han, Y., and Guo, W.: An improved multivariable integrated evaluation method and tool (MVIETool) v1. 0 for multimodel intercomparison, *Geosci. Model Dev.*, 14, 3079-3094, doi:10.5194/gmd-14-3079-2021, 2021.