

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

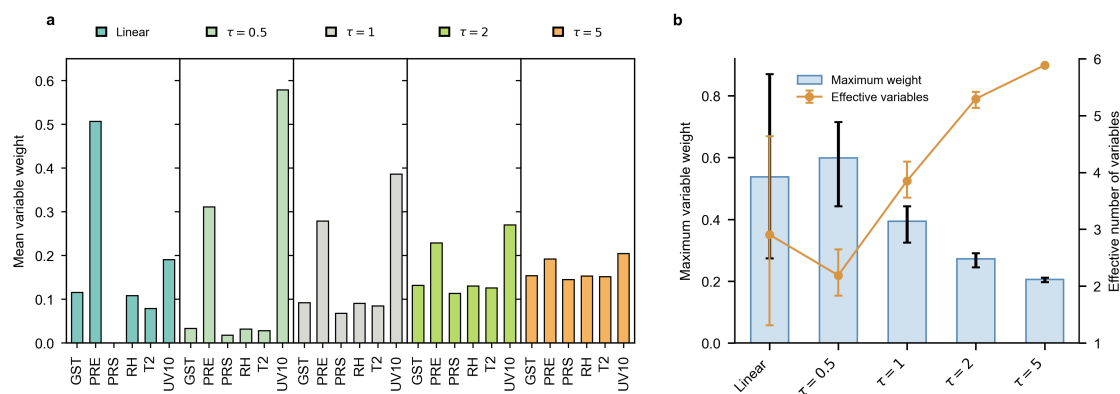


Figure S5. Sensitivity of the proposed variable weights to the normalization function. (a) Mean variable weights under linear normalization and temperature-scaled Softmax transformations. The original Softmax setting corresponds to $\tau=1$. (b) Weight-concentration diagnostics, including the maximum 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	0.103
	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
	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
CORR	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

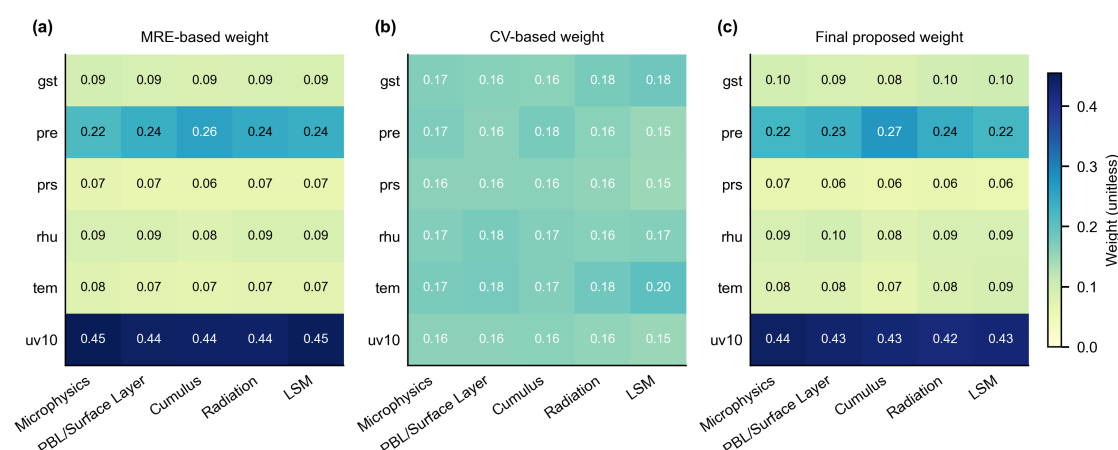


Figure 2. Weight distributions of the variables associated with the different physical processes of the WRF simulation in the arid region of Northwest. (a) MRE-based weight, (b) CV-based weight, and (c) final proposed weight. Weights are unitless.

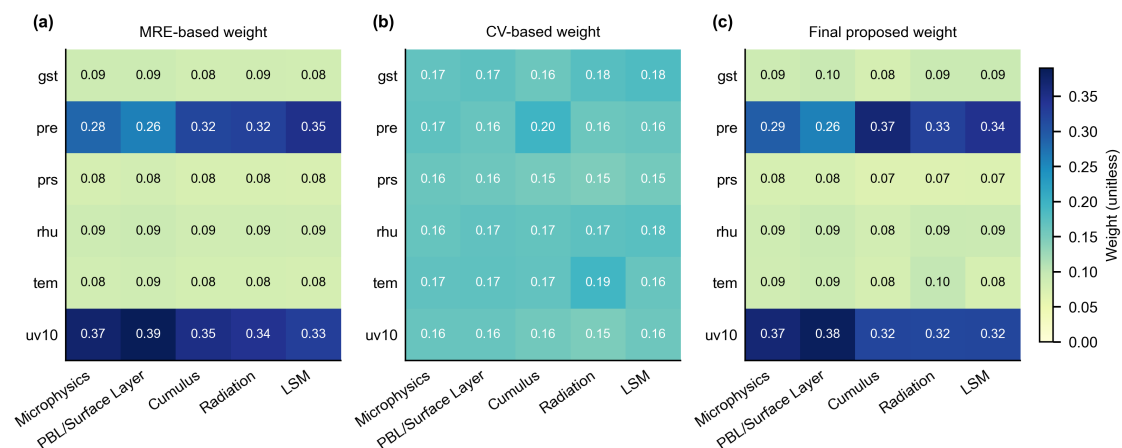


Figure 5. Weight distributions of the variables in the WRF simulations of the different physical processes in the coastal region of Southeast. (a) MRE-based weight, (b) CV-based weight, and (c) final proposed weight. Weights are unitless.

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.