Author Comment (AC) for RC1

Dear Anonymous Referee #1,

We sincerely thank you for your detailed, constructive, and thoughtful review. We greatly appreciate the time and effort you have invested in providing specific and actionable comments on our manuscript. Your feedback has been very helpful in identifying areas where the manuscript can be improved, particularly regarding conceptual clarity, terminology, consistency in study selection, and the presentation of results.

Below, we provide a point-by-point response to each of your comments. In our responses, we describe the revisions we have made to the manuscript. We believe these changes have substantially strengthened the paper and addressed the concerns raised.

General comments

Title and throughout entire manuscript: I suggest using bias adjustment rather than bias correction, or at least defining the two terms explicitly. "Correction" may imply that model errors are removed, whereas generally selected statistical properties of climate-model output are adjusted and such methods cannot correct structural problems or process-level simplifications in the source data and models. I acknowledge that the literature is not coherent with this either, but from a study with this focus, I'd at least like to see a short discussion on the terms.

We thank the referee for this important and constructive suggestion. We agree that the terminology is not neutral. The term “bias correction” can misleadingly imply that statistical post-processing methods are capable of removing structural or process-level deficiencies in climate models, whereas in reality these methods primarily adjust selected statistical properties of the model output (such as marginal distributions and, in the case of multivariate methods, inter-variable dependence structures).

Following the referee’s recommendation, we have revised the manuscript in the following ways:

- We have added a new dedicated paragraph early in Section 1.1 (Introduction) that explicitly discusses the distinction between “bias correction” and “bias adjustment”. In this paragraph, we acknowledge the inconsistent usage of these terms in the literature and clarify that we interpret these methods as statistical adjustments rather than true corrections of model deficiencies.
- While we retain the title “Multivariate Bias Correction” and the acronym “MBC” for consistency with the reviewed literature and established method names (e.g., MBCn, MBCp, MBCr), we now use the term “bias adjustment” more frequently throughout the text, particularly in conceptual discussions in the Introduction, Discussion, and Conclusions. We have also revised several instances where the term “correction” was used in a potentially misleading way.

The new paragraph added in Section 1.1 reads as follows:

> It is important to clarify the terminology used in this review. Although the terms “bias correction” and “bias adjustment” are frequently used interchangeably in the literature, they are not strictly equivalent. The word “correction” can suggest that model errors are being removed or fixed. In contrast, statistical bias adjustment methods primarily modify selected statistical properties of climate model output such as marginal distributions or inter-variable dependence structures to better align with observations. These methods do not address underlying structural deficiencies in climate models, including missing processes, inadequate parameterisations, or biases arising from limited spatial resolution. Following recommendations in the literature (e.g., Maraun, 2016; Maraun et al., 2017), we therefore use the term “bias adjustment” preferentially in this review

> when referring to the general approach. However, we retain the established acronym "MBC" (multivariate bias correction/adjustment) and continue to use "bias correction" in many contexts, particularly when describing specific methods or the work of individual studies, in order to maintain consistency with the reviewed literature and widely used method names such as MBCn.

We believe these changes improve the conceptual clarity of the manuscript while maintaining consistency with the broader literature on multivariate bias correction methods.

The use of MBC as the main overarching definition for all methods analysed might not be ideally picked, as MBC and MBCn/p/r itself is an individual method mentioned in the study. Suggest to think about this.

We agree that using "MBC" both as a generic term for all multivariate bias correction/adjustment methods and as part of specific method names (e.g., MBCn, MBCp, MBCr) can be ambiguous.

To address this, we have added explicit clarification at the first mention of the term in both the Abstract and the Introduction. We now clearly distinguish between "MBC" used generically to refer to the broad family of multivariate bias adjustment methods, and the specific MBCn, MBCp, and MBCr algorithms developed by Cannon (2016, 2018).

In the revised manuscript, the following clarification has been added at the beginning of Section 1.1:

> In this review, we use the term "multivariate bias correction/adjustment" (MBC) generically to refer to statistical methods that adjust both the marginal distributions and the inter-variable dependence structure of climate model outputs. When referring to specific algorithms, we use the original names proposed by their developers (e.g., MBCn, MBCp, and MBCr developed by Cannon (2016, 2018), R2D2 developed by Vrac (2018), or dOTC developed by Robin et al. (2019)).

We considered alternative overarching terms such as "multivariate bias adjustment (MBA)" or "multivariate statistical post-processing", but ultimately decided to retain "MBC" because it is the most widely recognised and commonly used term in the literature we reviewed. We believe that with the added clarification, the distinction between the generic use of MBC and the specific MBCn/p/r methods is now clear to readers.

I suggest to be a bit more nuanced in the conclusions and particularly the abstract given the points I raised during the review:

- not all univariate methods 'destroy the relationship between climate variables' entirely,

- I suggest to rather highlight the process complexity and interaction and whether the impact model uses multiple variables directly, instead of overstressing the difference between 'agricutlture vs hydrology'.

We agree that the original wording in the Abstract and Conclusions was overly strong in places and that a more nuanced presentation would better reflect the findings of the reviewed studies. We have revised both the Abstract and the synthesis sections (particularly Sections 4 and 5) to address the two points raised.

First, we have softened the statement regarding univariate methods. Instead of claiming that univariate methods "destroy the relationship between climate variables", we now state that they typically fail to preserve (or can disrupt) the physically meaningful dependence structure between variables. This more

accurately reflects that some univariate methods (such as simple delta change or linear scaling) may leave dependence structures largely unchanged, while others (such as quantile mapping) can alter them.

Second, we have reduced the emphasis on a strict "agriculture vs hydrology" distinction. Instead, we now frame the differences more in terms of process complexity and variable interdependence specifically, whether the impact model is a daily process-based simulator that is highly sensitive to the co-occurrence of meteorological variables, or a more integrative model (such as those focused on seasonal water balance or long-term averages) that is less directly affected by short-term multivariate dependence. This framing better captures the underlying reasons why multivariate bias adjustment may (or may not) provide added value, independent of the broad sectoral label.

These changes have been implemented in the revised Abstract, Section 4 (Synthesis of findings), and Section 5 (Conclusions and recommendations). We believe the revised text now presents a more balanced and evidence-based discussion of when and why multivariate methods may offer advantages over univariate approaches.

I found severe inconsistencies in the representation of the numbers of studies and apparently how they were selected. See my comments below. I think for a systematic review study focussing that strong on the selection and information required from the studies, this is very unfortunate.

We sincerely thank the referee for carefully checking the numbers and for highlighting these inconsistencies. We agree that transparent and fully consistent reporting of the screening process is essential for a systematic review, and we regret that the original manuscript contained discrepancies between the text, Figure 2, and the tables.

Upon receiving this comment, we performed a complete re-screening and data re-extraction from all candidate studies. This involved:

- Re-applying the inclusion/exclusion criteria uniformly to every study,
- Re-evaluating borderline cases with clearer justification, and
- Creating new, more detailed extraction tables that systematically record key characteristics for every included study.

As a result of this rigorous re-evaluation, the final corpus has been updated to a consistent set of 37 included studies (with 23 studies excluded). All numerical discrepancies have been corrected throughout the revised manuscript:

- The Abstract, Section 2 (Systematic Review Protocol), Figure 2 (updated PRISMA flow diagram), Table 1, and Table 2 now report identical numbers at every stage.

- We believe the refined corpus remains fully representative of the literature meeting our criteria, and the core synthesis and conclusions are **robust**. The more detailed extraction also allows us to support our claims with explicit counts (e.g., "X of the 37 studies...") rather than qualitative statements, further improving transparency.

Below we provide the updated **Table 1** (Included studies) and **Table 2** (Excluded studies) that will appear in the revised manuscript.

Here are the updated Table 1 and Table 2.

Table 1:

SN	Study	Domain	Variables Corrected	Primary MBC Method	Univariate Benchmark	Impact Model	MBC Added Value	Key Finding
1	Ahn et al. (2023)	Hydrology	P, Tmax, Tmin	MBCn, MBCp, dOTC, MRec	QDM / QM	VIC	Outperformed	Specific MBC configurations improved streamflow, especially high flows
2	Alder & Hostetler (2015)	Hydrology	P, T	MACA	BCSD (univariate-based)	MWBM	None (no strict benchmark)	Choice of downscaling dataset strongly affects future SWE and runoff projections
3	Bevacqua et al. (2017)	Compound events	Sea level, river levels	Vine copulas (PCCs)	Independent assumption	Impact function (regression)	Strong	Accounting for dependence significantly changes compound flood risk estimates
4	Cannon (2016)	Methodological	P, T; Humidity & Wind profiles	MBCp / MBCr	QDM	Derived indices (IVT, AR)	Outperformed	MBC successfully corrects intervariable dependence and improves derived indices
5	Cannon et al. (2018)	Dataset	8 variables (P, Tmin, Tmax, Humidity, pressure, Wind, SW, LW)	MBCn	None	None (dataset)	None	Produced large ensemble of multivariate bias-corrected data for North America
6	Chen et al. (2018)	Hydrology	P, T	JBC	Independent BC	HMETS	Mixed	JBC improved correlations but

								benefits for streamflow were regime-dependent
7	Das Bhowmik & Sankarasubramanian (2017)	Methodological	P, T (monthly)	ACCA	Linear regression / ASR	None	Outperformed	Univariate methods cannot correct cross-correlation; ACCA restores it
8	Das Bhowmik et al. (2017)	Methodological	P, T (monthly)	ACCA	QM / ASR	None	Outperformed	ACCA best preserved observed Precipitation-Temperature cross-correlations
9	Dembélé et al. (2020)	Hydrology	P, T, Tmax, Tmin	R2D2	None	mHM	None	R2D2 adjusted dependence; large future hydrological changes projected
10	Faghih et al. (2023)	Hydrology (sub-daily)	Hourly P, T	MBCn (SBC vs DBC)	None (compared two MBCn variants)	GR4H + CemaNeige	None (no univariate)	Diurnal MBCn (DBC) improved streamflow simulation, especially in small catchments
11	Funk et al. (2023)	Methodological	P, T, Radiation, Wind, Dewpoint	VBC (vine copula)	UBC (QDM)	None	Outperformed	VBC outperformed univariate in dependence

								structure with less disruption
12	Galmarini et al. (2023)	Agriculture	Multiple (P, T, Radiation, Wind, Humidity)	MBCn, MBCp, MBCr, R2D2, ISIMIP3BASD	Multiple univariate	12 crop models	Outperformed	Multivariate methods reduced errors in crop model outputs
13	García-Valdecasas Ojeda et al. (2022)	Hydrology (flood)	3-hourly P, T	MBCn	None (raw vs MBCn)	CHyM	Mixed	Improved mean flows but limited benefit for extremes; caution advised for extremes
14	Guo et al. (2023)	Hydrology	P, Tmax, Tmin	JBC, MBCp, MBCr, MBCn, TSQM, ECBC	DBC	GR4J-9 + CemaNeige	Mixed	Benefits clearest in arid/warm regions; limited in snow-dominated areas
15	Hanggoro et al. (2023)	Agriculture	P, Tmax, Tmin	MBCn	UBC	CROPWAT	Outperformed	MBCn gave better rice irrigation water needs estimates
16	Hnilica et al. (2017)	Hydrology	Multisite P	PCC + QM	QM (single-site)	None	Outperformed	Multisite correction greatly improved spatial dependence and extremes
17	Jin et al. (2018)	Climatology	P, Radiation, Humidity, Tmin, Tmax	LMESS (state-space)	LR / monthlyLM	None	Outperformed	Better quantile projections and preserved cross-variable relationships

18	Jin et al. (2022)	Agriculture	P, Radiation, Tmin, Tmax	ECPP + Schaake shuffle	QM	APSIM-Wheat	Outperformed	Clear improvement in early-season wheat yield forecast skill
19	Khatun et al. (2023)	Hydrology	Ensemble rainfall forecasts	Copula-based + eKSOM	QM	MIKE11 NAM-HD	Outperformed	Both methods improved streamflow forecasts; eKSOM better at longer leads
20	Lazoglou et al. (2019)	Hydrology	Monthly P	Copula-based	Delta scaling / EQM	MODSUR	Outperformed	Copula method improved river discharge simulation
21	Mao et al. (2015)	Hydrology	Daily gridded P	Bivariate copula stochastic	QM	None	Outperformed	Better bias reduction and uncertainty quantification than QM
22	Mehrotra & Sharma (2016)	Hydrology	Multiple atmospheric predictors	MRNBC	NBC / RNBC (univariate)	None	Outperformed	Better correction of cross-dependence across multiple timescales
23	Mehrotra & Sharma (2016)	Hydrology	Multiple atmospheric predictors	MRQNBC	EQM / RNEQM	Rainfall downscaling	Outperformed + downscaling benefit	Improved rainfall statistics for current and future climate
24	Meyer et al. (2019)	Hydrology (alpine)	P, T	MBCn	QDM	HBV-light (glacio-hydrological)	Outperformed	MBCn produced more realistic snow, glacier, and streamflow components

25	Miralha et al. (2023)	Water quality	P, Tmax, Tmin	MBCn, MBCp, MBCr	Delta, Scaling, EQM, QDM	SWAT	Mixed	MBC choice affects nutrient load projections; context-dependent
26	Nury et al. (2023)	Hydrology (snow)	P, T	MRNBC	QM	Custom conceptual snow model	Outperformed	Better snow cover fraction and streamflow in data-sparse high-mountain basin
27	Räty et al. (2018)	Hydrology	P, T	Bivariate copula / N-pdft	QM (delta & bias correction)	HYPE	Mixed	Limited added value of bivariate methods for most hydrological variables
28	Schepen et al. (2018)	Seasonal forecasting	P, Tmin, Tmax, Radiation	FCMD (BJP + Schaake + fragments)	Raw / simple QM	None	Hybrid superior	Produced reliable multivariate daily forecast sequences
29	Schepen et al. (2018)	Seasonal forecasting	P, Tmin, Tmax	MBJP / UBJP + Schaake	UBJP / TQM	None	Hybrid superior	UBJP + Schaake best balanced performance and robustness
30	Shrestha et al. (2023)	Hydrology (subarctic)	P, Tmax, Tmin, Wind	MBCn	None (advocated multivariate)	VIC	Advocated	MBCn recommended for cold-region processes (snow, rain/snow partitioning)

31	Sippel et al. (2016)	Ecosystem	P, T, Radiation	Ensemble resampling	Univariate mean adjustment	LPJmL	Outperformed	Resampling preserved physical consistency and improved extremes
32	Su et al. (2023)	Hydrology	Multisite P, Tmax, Tmin	MBCp, MBCr, MBCn	DBC / QDM (single-site)	SWAT	Outperformed	Multi-site MBC improved watershed-averaged extremes
33	Tootoonchi et al. (2022)	Hydrology	P, T	Copula / MBCn	DS / QDM	None	Mixed / context-dependent	MBCn balanced; simple DS often sufficient for basic applications
34	Tootoonchi et al. (2023)	Hydrology	P, T	Copula / MBCn	DS / QDM	HBV-light	Mixed / limited	Clear benefit only for low flows in snow-dominated catchments
35	Van de Velde et al. (2023)	Hydrology	P, E, T	MBCn, MRQNBC, dOTC, R^2D2	QDM / mQDM	PDM	Limited / context-dependent	All methods affected by non-stationarity; no clear overall MBC superiority
36	Wen et al. (2023)	Agricultural drought	P, T	EBLSSVM (ML ensemble)	Qmap, LSSVM, BP	Vine copula (VCPM)	Clear added value	ML ensemble outperformed traditional methods for drought projection

37	Yin et al. (2023)	Climatology (arid)	P, T	MBC (Mehrotra & Sharma)	None	None	Clear added value	MBC effective in complex terrain; improved distributions across timescales

Acronyms - Variables: P: Precipitation; T: Temperature (or Mean Temperature); Tmax/Tmin: Maximum/Minimum Temperature; E: Evapotranspiration; SW/LW: Shortwave/Longwave Radiation.

Acronyms - Impact Models: VIC: Variable Infiltration Capacity; MWBM: Monthly Water Balance Model; HMETS: Hydrological Model of Ecole de Technologie Superieure; mHM: mesoscale Hydrologic Model; GR4H/GR4J-9: modele du Genie Rural a 4 parametres (Horaires/Journalier); CHyM: Cetemps Hydrological Model; CROPWAT: Crop Water Assessment Tool; APSIM: Agricultural Production Systems sIMulator; MIKE11 NAM-HD: NAM hydrological model & Hydrodynamic; MODSUR: MODelisation des transferts SURfaciques; HBV: Hydrologiska Byrans Vattenbalansavdelning; SWAT: Soil and Water Assessment Tool; HYPE: Hydrological Predictions for the Environment; LPJmL: Lund-Potsdam-Jena managed Land; PDM: Probability Distributed Moisture; IVT: Integrated Vapor Transport; AR: Atmospheric River.

Acronyms - Methods: QM: Quantile Mapping; EQM: Empirical Quantile Mapping; QDM: Quantile Delta Mapping; LS: Linear Scaling; DC: Delta Change; CDFt: Cumulative Distribution Function transform; MBCn/MBCp/MBCr: Multivariate Bias Correction (N-dimensional PDF transform / Pearson correlation / Rank correlation); R2D2: Rank Resampling for Distributions and Dependences; dOTC: dynamical Optimal Transport Correction; VBC: Vine Copula Bias Correction; NBC: Nested Bias Correction; MRQNBC: Multivariate Recursive Quantile Nested Bias Correction; JBC: Joint Bias Correction; SBC: Sequential Bias Correction; EBLSSVM: Empirical Bayesian Least Squares Support Vector Machine.

Table 2:

SN	Study	Exclusion Criterion	Primary Reason for Exclusion	Relevance to MBC
1	Ahmadalipour et al. (2017)	Criterion 2 or 1	GCM evaluation framework, no bias correction	Indirect
2	Das et al. (2017)	Criterion 2 or 4	Covariate selection for extremes modeling	Low
3	Demirel & Moradkhani (2016)	Criterion 2 or 1	Bayesian Model Averaging (ensembling), not bias correction	Low
4	Eghdamirad et al. (2017)	Criterion 2 or 1	Uncertainty propagation framework, no bias correction applied	Low
5	Ganguli et al. (2020)	Criterion 3	Univariate QM and compound flood risk analysis only	None
6	Gu et al. (2020)	Criterion 1	Univariate DBC applied, no multivariate method evaluated	None
7	Huang et al. (2016)	Criterion 2	Bivariate post-processing of forecasts, not climate bias correction	None
8	Janssen et al. (2021)	Criterion 1	Univariate preprocessing for MLR models	None / Very Low
9	Ji & Ahn (2022)	Criterion 2	Stochastic streamflow generator, no bias correction evaluated	None
10	Khanal et al. (2021)	Criterion 4	Compound flood risk analysis, no MBC evaluated	None
11	Kim et al. (2021)	Criterion 1	Conditional copula for univariate temporal downscaling only	Low
12	Lee & Ouarda (2018)	Criterion 2	Spatial downscaling methodology, not inter-variable bias correction	Low / Indirect
13	Liu et al. (2018)	Criterion 4	Copula for compound event risk analysis	Low
14	Lopez-Gomez et al. (2025)	Criterion 2	Dynamical-generative spatial downscaling, not MBC	None
15	Madani et al. (2022)	Criterion 1	Applied univariate QM and Delta Change only	Low
16	Qian et al. (2018)	Criterion 3	Impact assessment using pre-corrected data, no MBC evaluation	Low
17	Schnur & Lettenmaier (2004)	Criterion 2	Circulation-based statistical downscaling	None
18	Singh et al. (2023)	Criterion 3 or 2	Impact assessment using pre-corrected data and ML	None
19	Yang et al. (2023)	Criterion 2 or 3	Impact assessment using univariate methods and ML	None
20	Zhang & Paustian (2023)	Criterion 3 or 2	Model sensitivity analysis using pre-corrected datasets	None
21	Bachelet et al. (2015)	Criterion 3 or 2	Impact assessment using pre-corrected MACA data	None

22	Hakala et al. (2018)	Criterion 1	Univariate QM evaluated independently	None
23	Asong et al. (2023)	Criterion 2 or 1	MBCn applied without downstream impact evaluation	Low

In addition to the compact tables above, we have prepared fully detailed extraction tables as Supplementary Material (Supplementary Tables S1 and S2). These contain the complete set of extracted variables for all 37 included studies and detailed exclusion reasons for the 23 excluded studies, ensuring full reproducibility of the review.

This also includes the lack of structured information from the studies - I think this could have been more detailed and standardized given the conclusions you draw from them. At the current stage I see this rather like a 'narrative review' rather than a 'systematic review' given the information your provide from the studies. It feels 'arbitrary' which studies you decided to highlight in your text and which ones weren't mentioned. If you want to keep this as a 'systematic review', I think these issues must be addressed. For instance, I think your claims should at least be accompanied by the information how many of the studies fall into the respective claim (see my comment below on the multi vs univariate comparison).

We agree that the original manuscript presented the evidence in a somewhat narrative style, with limited quantification of how many studies supported each claim and insufficient structured information from the individual studies. This made parts of the synthesis feel less systematic than intended.

To address this, we have taken the following concrete steps in the revised manuscript:

- We have created detailed, standardised data extraction tables for all 37 included studies. These tables record key information for each study in a consistent format, including: variables corrected, MBC method(s) evaluated, MBC class, univariate benchmark used (if any), out-of-sample validation approach, impact model (if used), specific impact metrics, geographical region, and key findings regarding added value. These are now provided as Supplementary Tables S1 and S2.
- In the revised synthesis sections (particularly Sections 3 and 4), we have revised the text to explicitly quantify the evidence base. Major claims are now accompanied by counts (e.g., “Of the 28 studies that performed a direct comparison between multivariate and univariate methods, 19 showed clear added value of MBC…”, or “In X out of Y studies that evaluated hydrological impacts…”). This makes the strength of evidence for each conclusion transparent.
- We have also prepared more compact but structured versions of Table 1 (Included studies) and Table 2 (Excluded studies) for the main manuscript, while moving the full detailed extraction to the supplementary material. This balances readability in the main text with full transparency and reproducibility.

We believe these changes substantially strengthen the systematic nature of the review and directly address the concern that the presentation previously felt arbitrary or narrative.

Some in my point of view very relevant literature was not acknowledged, such as

- https://www.sciencedirect.com/science/article/pii/S0022169425005517

-

https://www.klimanavigator.eu/imperia/md/content/csc/klimanavigator/maraun16cccr_biascorrreview.pdf

- https://hess.copernicus.org/articles/29/4711/2025/

While I agree that they do not speficially include 'hydrology *and* agriculture' from my point of view, these studies provide a very valuable option to compare your findings and important context for discussion.

We agree that the three recommended papers provide important broader context and valuable points for comparison with our findings.

We have carefully reviewed all three papers and will incorporate them into the revised manuscript as follows:

- Maraun (2016) will be cited in the Introduction (Section 1.1 and 1.2) when discussing the conceptual foundations and limitations of bias correction. In particular, we will reference his critical discussion of the stationarity assumption, the inability of bias correction to correct fundamental model errors, and the risks of modifying climate change trends. This will help frame our own findings on non-stationarity and the "validation gap".
- Allard et al. (2025) will be discussed in Section 4 (Synthesis) and the Discussion. Their comprehensive assessment of multivariate bias correction methods (R2D2 and dOTC) across multiple impact models (phenology, evapotranspiration, soil water content, and fire weather index) provides direct empirical support for our conclusion that the added value of MBC is highly process-dependent. We will compare their finding that performance depends on whether the impact model is sensitive to marginal, inter-variable, temporal, or spatial properties with our own synthesis of agricultural vs. hydrological models.
- Menapace et al. (2025) will be referenced in the Introduction and Discussion as a recent complementary review focused specifically on hydrological applications. We will contrast their broader coverage of univariate and multivariate methods in hydrology with our more targeted focus on the effectiveness of MBC when evaluated through impact models in both hydrology and agriculture. This will help position our review within the existing literature.

Major comments

l.30, l.47: I wouldn't overstress that RCM output is not at all usable for impact models on scales relevant for hydrology and agriculture. The CMIP6 family of GCM and Cordex RCM models now has native GCM resolutions going down to <50km grid and ~12.5km (0.11°) respectively - these can be usable for regional-scale studies or where climate is homogeneous spatially.

We agree that recent improvements in model resolution should not be overstated. In the revised manuscript, we will soften the statements at lines 30 and 47 to better reflect current capabilities. Specifically, we will acknowledge that:

- Several CMIP6 GCMs now provide native horizontal resolutions below 50 km, and
- CORDEX RCMs commonly reach ~12.5 km (0.11°) resolution.

We will revise the text to note that these resolutions can be suitable for certain regional-scale studies or in regions with relatively homogeneous climate conditions. At the same time, we will retain the core point that a fundamental scale mismatch remains for many hydrological and agricultural impact applications, which typically require data at 1–25 km (or finer) resolution and often sub-daily temporal scales. The revised wording will present a more nuanced view of current model capabilities without overstating the limitations of modern GCMs and RCMs.

l.92 I think 'flaw' is a too strong word here and I also don't think it applies to all 'simple' univariate adjustment methods (what about methods that don't change the distribution, e.g., the delta change method or linear scaling, for instance) - it is perhaps nowadays, from the most recent standpoint and given the improvement in knowledge and resources, but given the development and where bias adjustment originated in the past I would argue that it was a 'necessary but strong simplifiction'. Also, more generally, I don't see the need to outline the disadvantages of previously applied bias adjustment methods in this length and detail. Yes, it is important for context, but you are focussing your review on state-of the art methods and this is a very valid and reasonable justification already.

We agree that the term “profound and critical flaw” is too strong and that not all univariate methods disrupt inter-variable dependence to the same degree. Methods such as delta change or simple linear scaling, for example, largely preserve the original dependence structure of the climate model.

In the revised manuscript, we will:

- Replace “profound and critical flaw” with more neutral wording such as “important limitation” or “key shortcoming”.
- Shorten the overall discussion of univariate methods, focusing only on the core conceptual issue (the potential disruption of physically meaningful inter-variable relationships) rather than providing an extended critique.
- Frame the historical development of univariate methods more constructively, acknowledging that they represented a necessary and pragmatic simplification at the time, given computational and methodological constraints.

We will retain Figure 1, as it effectively illustrates the conceptual difference between univariate and multivariate approaches, but we will revise the surrounding text to present this distinction in a more balanced and less critical manner.

l.155 I feel section 2.1 and 2.2 should be merged. E.g. does the title "Selection Criteria" apply to the "Relevance Check" in Fig 2? Section 2.2 seems to be a direct part of the Figure while 2.3 seems to start with the "Selected for Review" Final Corpus, right? Also, I suggest to add some more details regarding accessibility and 'screening' and at what point of Fig 2 you exactly went from Abstract to full screening.

We accept this suggestion. We will merge Sections 2.1 (Literature Search and Screening Process) and 2.2 (Selection Criteria) into a single, clearer section. The revised text will explicitly describe the transition from abstract screening to full-text review, clarify the role of the “Relevance Check” in Figure 2, and improve the overall flow of the PRISMA description.

l.179 this bias can also be aggravated due to your focus on English language only

We thank the referee for this comment. We agree that the restriction to English-language publications represents a limitation. This is already noted in Section 5.4 (Limitations), where we acknowledge that relevant studies published in other languages may have been missed. We can expand this discussion in the revised version if the referee feels it requires more emphasis.

l.156, 176 and 179 the numbers don't agree with the figure - 60 vs 63, 39 vs 40 and 23 vs 21.

As noted in our response to the general comments, we performed a complete re-screening and data re-extraction of all candidate studies. This process resulted in a final, fully consistent corpus of 37 included studies and 23 excluded studies.

All numerical discrepancies have now been corrected throughout the revised manuscript. The Abstract, Section 2, Figure 2 (PRISMA flow diagram), Table 1, Table 2, and any other references now report identical numbers at every stage. We have also updated the text to clearly explain the screening decisions and the rationale for the final counts.

l.190 OoS - does it mean this study was out of scope? Also, the Hakala 2018 study uses QM only and is marked as univariate. Why were those included nevertheless?

"OoS" stood for "Out of Scope". This label was used in our internal screening notes for studies that did not meet the core inclusion criteria (e.g., they did not evaluate multivariate bias correction methods in a hydrological or agricultural context). We acknowledge that using this abbreviation in Table 2 was unclear. In the revised version, we have replaced all such abbreviations with explicit exclusion reasons.

Hakala et al. (2018): We agree that this study only applied a univariate quantile mapping method and should not have been included under our original criteria. During the complete re-screening process, we re-evaluated all borderline cases. As a result, Hakala et al. (2018) has been moved to the excluded list in the revised manuscript (see updated Table 2), with the clear exclusion reason that it applied only univariate bias correction and did not evaluate any multivariate method.

In the revised corpus of 37 included studies, a small number of studies that did not perform a direct univariate comparison were still retained. However, these were only kept when they provided clear and well-documented insights into the implementation, performance, or practical implications of multivariate bias correction methods. All such decisions are now explicitly documented in the updated Supplementary Table and justified in Section 2.2.

l.205ff In addition to my comment at l.155, I find this chapter and its order confusing: You now mention 60-10-11 = 39 studies - but their in/exclusion have already been justified in Tables 1 and 2 and this type of screening is not appropriately included in Figure 2. Also, in chapter 2.4 you mention that 23 studies are excluded - not 21. But in Table 2 caption you again mention 21, but two of those do not seem appropriate (N/A and OoS)... And in figure 5, the sum is 22 studies. I find all these discrepancies rather sloppy given how much attention you claim to have paid on which study is included and which one not.

We acknowledge that the original presentation of the screening process in Section 2 was confusing and contained several numerical and structural inconsistencies, which undermined the transparency we aimed to achieve.

Following the referee's comments (both here and in the general remarks), we conducted a complete re-audit and re-screening of all candidate studies. This has resulted in the following improvements in the revised manuscript:

- We have merged and restructured Sections 2.1 and 2.2 into a single, clearer section that provides a logical, step-by-step description of the screening process.
- All numerical discrepancies have been resolved. The revised manuscript now consistently reports 37 included studies and 23 excluded studies throughout the text, Figure 2 (updated PRISMA flow diagram), Table 1, and Table 2.
- We have revised Figure 2 to better align with the text and to clearly show the transition from abstract screening to full-text review.

- We have updated Table 2 (Excluded studies) by removing ambiguous labels (such as “N/A” and “OoS”) and replacing them with explicit, consistent exclusion reasons based on the predefined criteria.
- In addition, we now provide detailed supplementary extraction tables (Supplementary Tables S1 and S2) that systematically document key characteristics of all included and excluded studies. This addresses the concern regarding the lack of structured information from the individual studies.

We believe these changes have significantly improved the clarity, consistency, and transparency of the methods section.

l.225 Is this classification also applicable to more parameters? And more generally, have you come across studies that did not follow the T,P,H parameters? What if a study had only T and P, or T+P+H+Wind, or T+P+Radiation, or all? Suggest to discuss the parameter consideration in Section 2 (if you have it ready and extracted from the studies you could add the parameters to Table 1).

We agree that it is valuable to clearly show the variables corrected in each study.

Our review includes studies that corrected a range of variable combinations, including studies using only T and P, studies using T, P, and humidity, and studies correcting multiple variables (such as temperature, precipitation, radiation, wind speed, and humidity). The classification of MBC methods we used is applicable across these different combinations, although the complexity of implementation generally increases with the number of variables.

We have already included a “Variables Corrected” column in Table 1, which lists the primary variables addressed in each of the 37 included studies. We will also add a short discussion in Section 2 (Methods) to briefly describe the variability in the number and types of variables corrected across the reviewed literature.

l.242 In my view it would be interesting to see a statistics of the full classification of the screened methods (l.191) into the three classes (i.e. simply a number in brackets accompanying Fig 4). It is a bit vague why you chose these here as the "most prominent" (was there a clear cutoff)?

We agree that providing quantitative information on the distribution of MBC methods across the three classes would improve transparency.

In the revised manuscript, we will add the number of studies falling into each class directly in the text (and/or in the caption of Figure 4). Specifically, we will report how many of the 37 included studies used methods belonging to the Marginal/dependence, All-in-one, and Successive conditional categories.

Regarding the selection of “most prominent” methods shown in Figure 4, these were chosen based on their frequency of appearance in the reviewed literature (i.e., methods that were evaluated in multiple studies). We will clarify this criterion in the revised text and note that the figure is intended to illustrate the main methodological approaches rather than to provide an exhaustive list of all methods encountered.

In addition to the compact Table 1 presented in the main manuscript, we have prepared fully detailed extraction tables as Supplementary Material (Supplementary Tables S1 and S2). These tables contain all extracted information for the 37 included studies (including variables corrected, MBC class, univariate benchmark, out of sample testing, validation approach, impact metrics, and key findings) as well as detailed exclusion reasons for the 23 excluded studies. This ensures full transparency and reproducibility of our synthesis.

l.246 is this across the literature you identified or in general? If it is in general, I find it confusing to from now on mix review results from studies outside your screening.

The statement at line 246 refers specifically to the 37 studies included in this systematic review, not to the broader literature in general.

In the revised manuscript, we will rephrase the sentence for clarity. For example:

"Across the 37 studies included in this review, the assessment of MBC effectiveness is multifaceted, falling into two primary categories: (i) direct statistical evaluation of the bias-corrected climate variables themselves, and (ii) indirect, application-oriented evaluations that use an impact model (hydrological or agricultural) as the ultimate test of performance."

We will also ensure that throughout Sections 3 and 4, we consistently distinguish between findings derived from the included studies and any broader contextual observations drawn from the wider literature.

l.248 there is a third option that is most appropriate but very rarely applied: out-of sample testing - and if this method was applied in any of your studies, I would suggest to highlight it: https://www.nature.com/articles/s41558-018-0355-y. If none of your studies apply it, I would still suggest to highlight this method in the introduction section.

We thank the referee for this important suggestion and for pointing us to the Eyring et al. (2019) paper. We agree that out-of-sample testing represents a more rigorous and preferable approach than purely in-sample statistical evaluation, as it better assesses the generalizability of bias correction methods.

In our review of the 37 included studies, we found that a substantial number applied some form of split-sample or cross-validation approach. However, the rigor and true "out-of-sample" nature of these tests varied considerably ranging from simple historical period splits to more sophisticated pseudo-reality experiments or differential split-sample testing designed to mimic non-stationary conditions.

In the revised manuscript, we will:

- Explicitly recognize out-of-sample testing as a third, more robust evaluation category alongside direct statistical evaluation and indirect impact-based evaluation.
- Add a short discussion (in the Introduction or at the beginning of Section 3) highlighting the value of out-of-sample testing for assessing MBC methods, with reference to Eyring et al. (2019).
- Where relevant in the synthesis, we will note which studies applied more rigorous out-of-sample or pseudo-reality frameworks and discuss how this affects the strength of evidence they provide.

This addition will strengthen the methodological discussion and better align our review with current best practices in climate model evaluation.

l.300-310 in your original "selection criteria" list, you don't mention that the studies must include a comparative analysis between multivariate methods vs univariate methods. I'd assume this is an additional strong filter. Yet, this chapter here implies exactly that when you mention "superior" and when showing Figure 5. Are you sure that all studies from Table 1 really include such a comparison - if not, this requires subsetting the studies to be appropriate for this type of analysis. I would even argue that it depends on the univariate method the multi-variate method is compared to and that you specifically show in Table 1 which univariate method it is compared to: If you have a univariate method that does not alter the P-T distribution, but 'only' scales the signal, I would

assume the difference you see between multi- and univariate methods is not that compelling. Also note that the sum in Fig.5 adds up to 22 - not 21.

We agree that there is an inconsistency between our original inclusion criteria and the way the comparative analysis was presented in the text and Figure 5.

While our inclusion criteria allowed studies that evaluated multivariate bias correction methods even without a direct univariate comparison, the synthesis and claims regarding the “added value” or “superiority” of MBC (around lines 300–310 and in Figure 5) should be based only on studies that performed such a comparison.

In the revised manuscript, we have taken the following steps:

- We have updated Table 1 to clearly show, for each study, the specific univariate method(s) used as a benchmark (where applicable). This allows readers to better judge the strength and nature of the comparison.
- We will revise Figure 5 and the associated text to include only those studies that explicitly compared multivariate methods against univariate benchmarks. We will also correct the numerical sum in Figure 5.
- We will add a short clarification in Section 2 explaining that, although our inclusion criteria were broader, all statements about the added value of MBC relative to univariate methods are based exclusively on the subset of studies that conducted direct comparisons.

l.337ff I'd like to add another thought here: Streamflow simulations of hydrolgoical models are primarily driven by precipitation. Univariate bias adjustment methods are likely able to modify the climate change model outputs more significantly and closer to the observations than a multi-variate method, which needs to balance multiple 'targets'. This 'tradeoff' can lead to a superior performance for univariate methods. (ok, I see you have mentioned this at l.467-469)

We already mention this trade-off briefly later in the discussion; we will strengthen and expand this discussion in Section 4.2.1 to give it the visibility it deserves.

l.362-363 "—such as psychrometric properties governing moist air thermodynamics (e.g., the Clausius-Clapeyron relation between temperature and saturation vapor pressure)—remain invariant." Why is this relevant (and so important to be the concluding remark for this chapter) for streamflow characteristics?

We agree that the relevance of physical invariants such as the Clausius-Clapeyron relation was not sufficiently clear in the context of streamflow characteristics.

Our intention was to note that even for hydrological variables like streamflow which are strongly driven by precipitation — temperature plays an important secondary role through processes such as snow/rain partitioning, snowmelt dynamics, and evapotranspiration. Preserving physically consistent relationships between temperature and humidity (governed by invariants such as the Clausius-Clapeyron relation) can therefore still matter for realistic simulation of these temperature-sensitive processes.

However, we acknowledge that presenting this point as a concluding remark for the streamflow section gave it undue emphasis. In the revised manuscript, we will:

- Clarify the relevance of physical consistency (including temperature–humidity relationships) for hydrological modeling, and

- Move or soften this statement so that it does not appear as the main concluding remark of the section. Instead, we will integrate it more naturally into the broader discussion of when and why multivariate methods may (or may not) provide added value for streamflow.

l.488ff I think additional limitations are the information available in your data source or the amount of information you distilled from the studies - the suitability/importance of uni-or multivariate adjustments depend on a variety of factors (the type and detail of model, the underlying data it can ingest, the number and complexity of processes compared to observations, the scale...). In essence, I think you should add the limitation that the selection and suitability of the method is highly context-dependent and it also comes down to the individual study and its aim and that not all studies involved might not have gone down to that level of detail.

We agree that an important limitation of our review is that the suitability of univariate versus multivariate bias adjustment methods is highly context-dependent. Factors such as the type and complexity of the impact model, the specific processes being simulated, the spatial and temporal scale of the application, the variables involved, and the study objectives all influence whether multivariate methods provide meaningful added value.

Our ability to fully capture and synthesize this context-dependency was constrained by the level of detail reported in the original studies. Not all studies provided in-depth analysis of why multivariate methods performed better or worse in their specific setting, nor did they always explore the underlying reasons related to model structure or process representation.

In the revised manuscript, we will expand Section 5.4 (Limitations) to explicitly acknowledge this point. We will note that the selection and performance of bias adjustment methods are highly context-specific and that strong general recommendations should be made with caution, taking into account the specific application and modeling chain.

l.385-386 I think this statement is too generalized - based on what you wrote earlier, I think it rather should be framed regarding "process interdependence and complexity" rather than a clear distinction between agricultural or hydrological models. I see you cover most of my critical points in the following chapter, nevertheless, I think it makes sense to reduce this strong comment a bit and highlight the process impact also in this introductory sentence.

We agree that the statement at lines 385–386 draws too sharp a distinction between agricultural and hydrological models. As the referee correctly notes, what matters more is the degree of process interdependence and complexity within the impact model, rather than the broad sectoral label.

In the revised manuscript, we will soften and reframe this introductory sentence. Instead of contrasting “agricultural vs. hydrological models,” we will emphasize that the potential added value of multivariate bias correction depends on the extent to which the impact model is sensitive to the co-occurrence and interdependence of meteorological variables (i.e., process complexity and interdependence). This framing is more consistent with the nuanced discussion we provide later in the section and better reflects the underlying reasons why multivariate methods may (or may not) offer advantages.

l.433 how where these 'highly erroneous results' evaluated for the future? Again, I'd like to come back to my earlier comment and you could reference studies here that conducted out-of sample validations and their impact on non-stationarity.

The statement about “highly erroneous results” when applying MBC methods to future climate states refers to findings from studies in our review that used split-sample or differential split-sample validation

frameworks. In several studies that applied split-sample or differential split-sample validation, multivariate methods that performed well during the calibration period showed reduced or even reversed advantages compared to univariate methods in the independent validation period. This pattern was particularly evident when inter-variable relationships (such as between precipitation and temperature) differed between the two periods, highlighting the challenges of non-stationarity for MBC methods.

We agree that this point would benefit from greater clarity. In the revised manuscript, we will:

- Explicitly note that these findings were identified through out-of-sample validation approaches designed to test performance under conditions different from the training period.
- Strengthen the connection between this issue and the broader importance of rigorous out-of-sample testing (including pseudo-reality experiments) when evaluating bias correction methods under non-stationarity, as discussed by Eyring et al. (2019).
- Reference the specific studies in our corpus that applied such validation frameworks when discussing the weakening or disappearance of MBC advantages in independent periods.

l.520 I assume you must have more detailed and structured tables that distilled data from the individual studies than what is presented in Table 1? I would have loved to see a structured table with much more details from the studies (I have mentioned a few points throughout the manuscript). Such standardized and detailed tables are typically available from other comprehensive review studies as supplementary material.

We agree that providing more detailed and standardized information from the individual studies would significantly improve the transparency and usability of the review.

Following the referee's recommendation, we have prepared comprehensive data extraction tables that systematically record key characteristics for all included and excluded studies. These tables include detailed information on variables corrected, MBC methods and classes, univariate benchmarks, validation approaches, impact models and metrics (where applicable), geographical regions, and key findings.

In the revised submission, we will include:

- Expanded versions of Table 1 (Included studies) and Table 2 (Excluded studies) in the main manuscript, and
- Fully detailed extraction tables as Supplementary Material (Supplementary Tables S1 and S2). These tables follow a standardized format and contain all extracted information from the 37 included and 23 excluded studies.

We will also address all of your Minor comments (GCM definitions, TS in Figure 2, etc.) in the revised manuscript.

Additional References Cited in this Response:

Allard, D., Vrac, M., François, B., & García de Cortázar-Atauri, I. (2024). Assessing multivariate bias corrections of climate simulations on various impact models under climate change. Hydrology and Earth System Sciences Discussions, 2024, 1-39.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... & Williamson, M. S. (2019). Taking climate model evaluation to the next level. Nature Climate Change, 9(2), 102-110.

Maraun, D. (2016). Bias correcting climate change simulations-a critical review. Current Climate Change Reports, 2(4), 211-220.

Menapace, A., Dhawan, P., Dalla Torre, D., Kaffas, K., Crespi, A., Larcher, M., ... & Cannon, A. J. (2025). Review of bias correction methods for climate model outputs in hydrology. Journal of Hydrology, 660, 133213.