RC1:

This paper investigates when and why bivariate and multivariate Liang–Kleeman Information Flow (L-K IF) estimates diverge in land–atmosphere applications. The manuscript has clear potential and addresses an interesting methodological problem of practical relevance. However, substantial revisions are required to strengthen the manuscript.

We thank the reviewer for their thorough and constructive comments.

Below, we provide a point-by-point response to each comment. Reviewer comments are shown in black, our responses in blue, and the corresponding revisions in the manuscript are highlighted in yellow.

Reviewer's comments:

1. The paper defines ΔIF as an absolute difference in Section 2.2, but Appendix A later writes ΔIF expressions without absolute value signs.

Response: We thank the reviewer for pointing out this inconsistency. We agree there is an ambiguity in the original manuscript.

To resolve this, we have revised the manuscript to distinguish explicitly between the signed difference and its magnitude. Specifically, ΔIF is now defined as the signed difference between the bivariate and multivariate IF estimates, i.e., $\Delta IF = T^{bi} - T^{mv}$, while its absolute value, $|\Delta IF|$, is used to quantify the magnitude of divergence. We have applied this notation consistently throughout the manuscript, including Section 2.2, Appendix A, and all figures and captions, to avoid ambiguity.

2. The significance assessment of ΔIF via a paired two-tailed Student's t-test at each time step is not sufficiently justified. Because the IF estimates are time-varying, serially dependent, and derived from filtered time series, the independence assumptions behind a pointwise t-test may not hold.

Response: We thank the reviewer for this comment. We would like to clarify that, in our implementation, the paired two-tailed Student's t-test at each time step is not applied across temporally successive observations. Instead, at each monthly time step, the test is performed between the bivariate and multivariate IF estimates using the set of grid-cell values within the study region. Thus, the replication for the paired comparison is provided by the ensemble of grid cells at a given time step, rather than by the temporal dimension.

To evaluate the sensitivity of this significance assessment to the testing framework, we additionally applied paired bootstrap resampling across grid cells while preserving the pairing between the bivariate and multivariate IF estimates. The bootstrap-based confidence intervals identified the same significant periods as

the paired two-tailed Student's t-test, indicating that the inferred ΔIF significance pattern is robust to the choice between parametric and nonparametric interval estimation.

In addition, because significance was evaluated repeatedly across monthly time steps, we applied the Benjamini–Hochberg false discovery rate (FDR) procedure to the full set of monthly p-values to control the expected proportion of false discoveries arising from multiple testing. The FDR correction did not alter the temporal pattern of significant ΔIF, indicating that the main results are robust to multiple-testing adjustment. We have revised the Methods section accordingly and now note that, because neighboring grid cells may still exhibit spatial autocorrelation and the IF estimates are derived from temporally filtered series, the resulting confidence intervals should be interpreted as approximate regional uncertainty bounds. The added paragraph reads as follows:

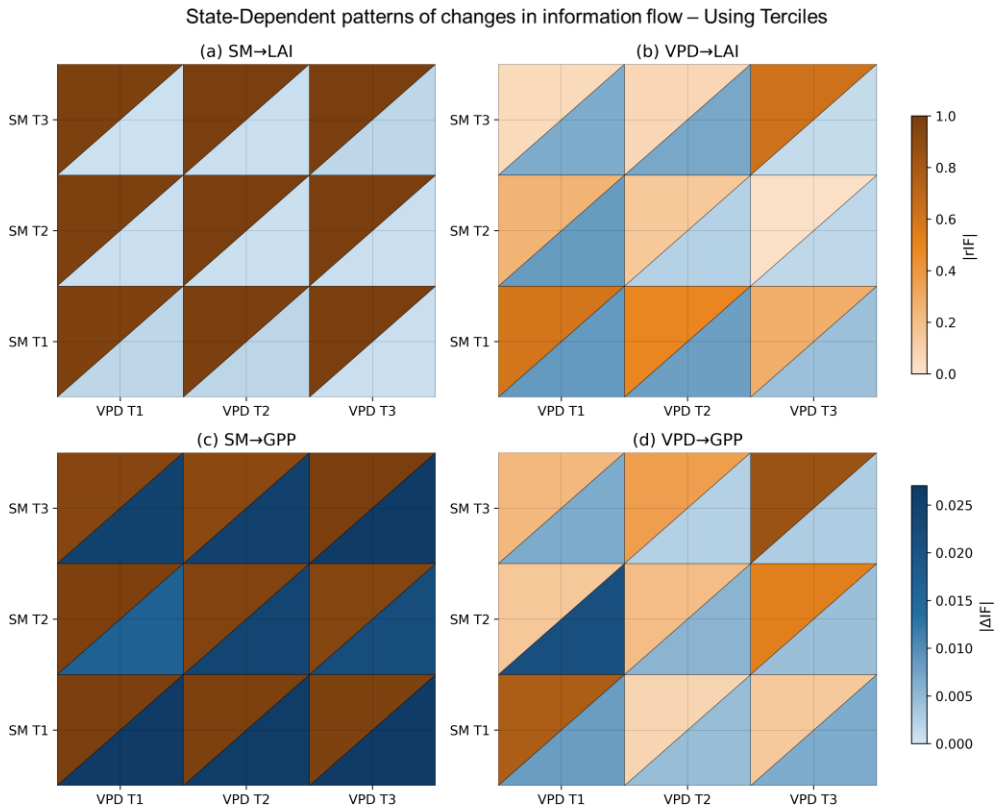
To evaluate the sensitivity of this significance assessment to the choice of testing framework, we additionally performed paired bootstrap resampling across grid cells at each monthly time step, while preserving the pairing between the bivariate and multivariate IF estimates. The bootstrap-based 95% confidence intervals identified the same significant periods as the paired t-test, indicating that the inferred ΔIF significance pattern is robust to the choice between parametric and nonparametric interval estimation. Because significance was assessed repeatedly across monthly time steps, we further applied the Benjamini–Hochberg false discovery rate (FDR) procedure to the resulting monthly p-values to control the expected proportion of false discoveries associated with multiple testing (Benjamini and Hochberg, 1995; Wilks, 2016). The FDR correction did not alter the temporal pattern of significant ΔIF.

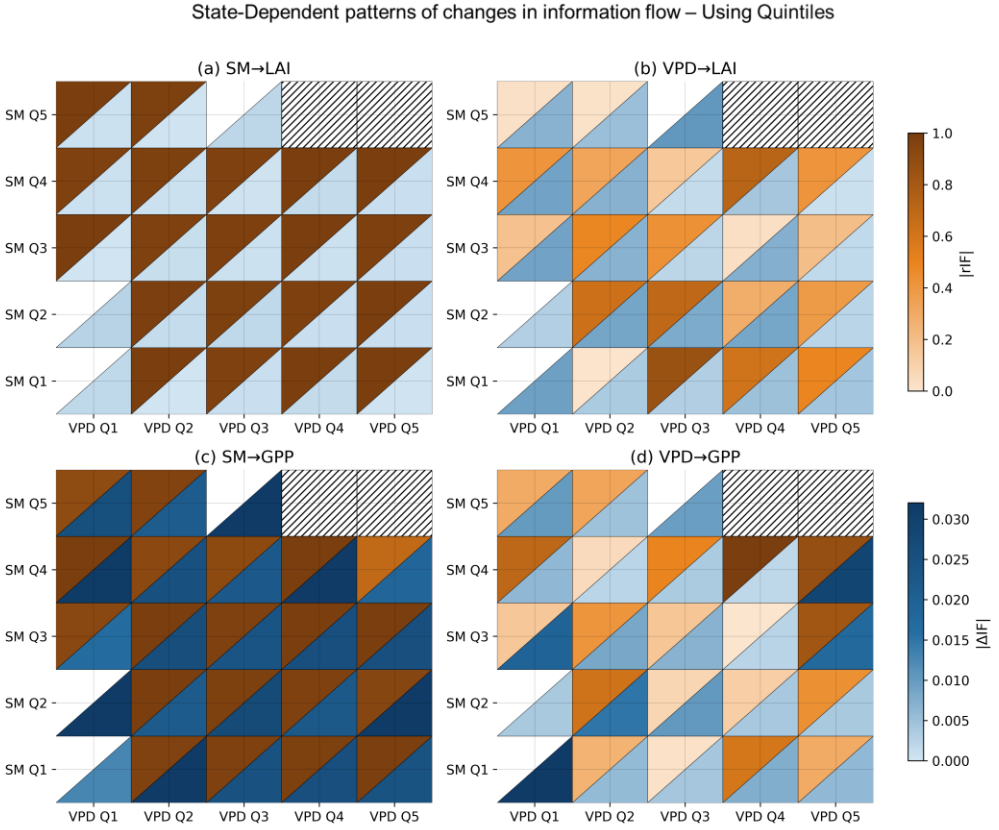
However, because the replication is provided by the spatial ensemble of grid cells rather than by temporally successive observations, and because neighboring grid cells may exhibit spatial autocorrelation while the IF estimates themselves are derived from temporally filtered series, these confidence intervals should be interpreted as approximate regional uncertainty bounds rather than exact pointwise significance tests.

3. The choice of four quantile-based regimes is reasonable but somewhat arbitrary. The manuscript should explain why quartiles were selected instead of terciles, quintiles, or physically defined thresholds. A short sensitivity test would be helpful.

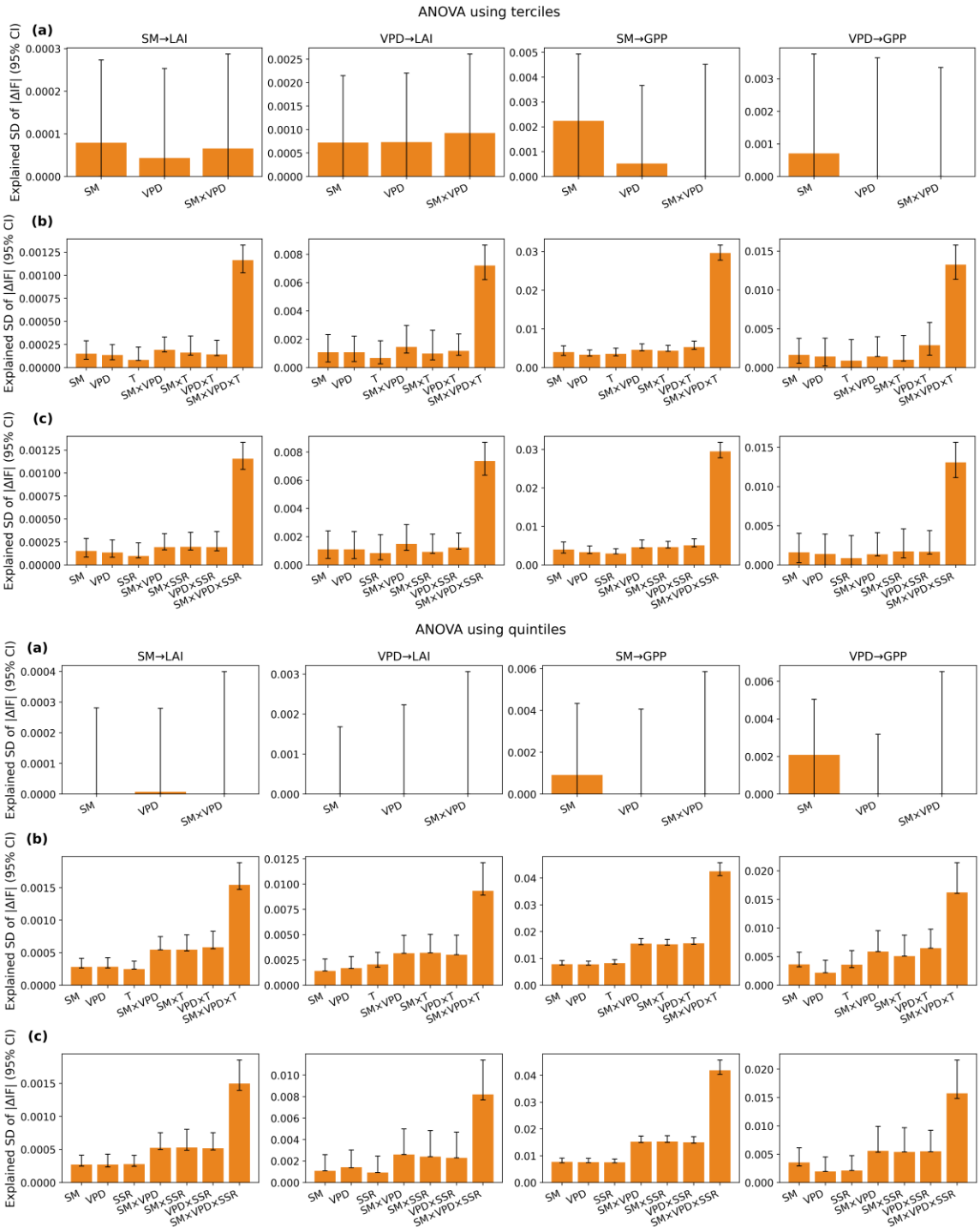
Response: We thank the reviewer for this helpful suggestion. We agree that the choice of quartiles is not unique and should be justified. In the revised manuscript, we now clarify that quartiles were selected as a practical compromise between hydroclimatic state resolution and sample support within each joint SM–VPD regime. Terciles provide a coarser partition and can mask intermediate state dependence, whereas quintiles offer finer resolution but increase sparsity in some regime combinations, leading to less stable regime-specific estimates.

As a sensitivity test, we repeated the regime-based analyses using tercile- and quintile-based classifications. For the split-cell |rIF|–|ΔIF| analysis, the main qualitative patterns remained unchanged across all three choices: SM-driven couplings remained comparatively stable, whereas VPD-driven couplings showed stronger regime dependence. For the ANOVA, the main conclusions were also robust, particularly the dominant role of higher-order interactions such as SM×VPD×T and SM×VPD×SSR, especially for the GPP pathways. Some weaker two-way components, particularly in VPD-driven pathways, showed greater sensitivity to the number of bins. We have added this clarification to the Methods and report the sensitivity analysis in the Supplementary Fig. S3 and S5.


State-Dependent patterns of changes in information flow – Using Terciles
(a) SM→LAI
(b) VPD→LAI
(c) SM→GPP
(d) VPD→GPP
SM T1
SM T2
SM T3
VPD T1
VPD T2
VPD T3
|rIF|
|ΔIF|


State-Dependent patterns of changes in information flow – Using Quintiles
(a) SM→LAI
(b) VPD→LAI
(c) SM→GPP
(d) VPD→GPP
SM Q1
SM Q2
SM Q3
SM Q4
SM Q5
VPD Q1
VPD Q2
VPD Q3
VPD Q4
VPD Q5
|rIF|
|ΔIF|

Supplementary Fig. S3. Differences (ΔIF) and correlations (rIF) between bivariate and multivariate Information Flow across SM and VPD terciles (Q1–Q3; top panel) and quintiles (Q1–Q5; bottom panel) for Southeast China. (a) SM→LAI, (b) VPD→LAI, (c) SM→GPP, and (d) VPD→GPP. Each cell is divided into an upper triangle (|rIF|: magnitude of Pearson correlation between bivariate and multivariate IF) and a lower triangle (|ΔIF|: absolute difference between bivariate and multivariate IF). Colour intensities are scaled independently for each triangle using separable colormaps. Cross-hatched cells indicate combinations with no data. (SM= Soil Moisture, VPD= Vapor Pressure Deficit, LAI= Leaf Area Index, GPP= Gross Primary Production).

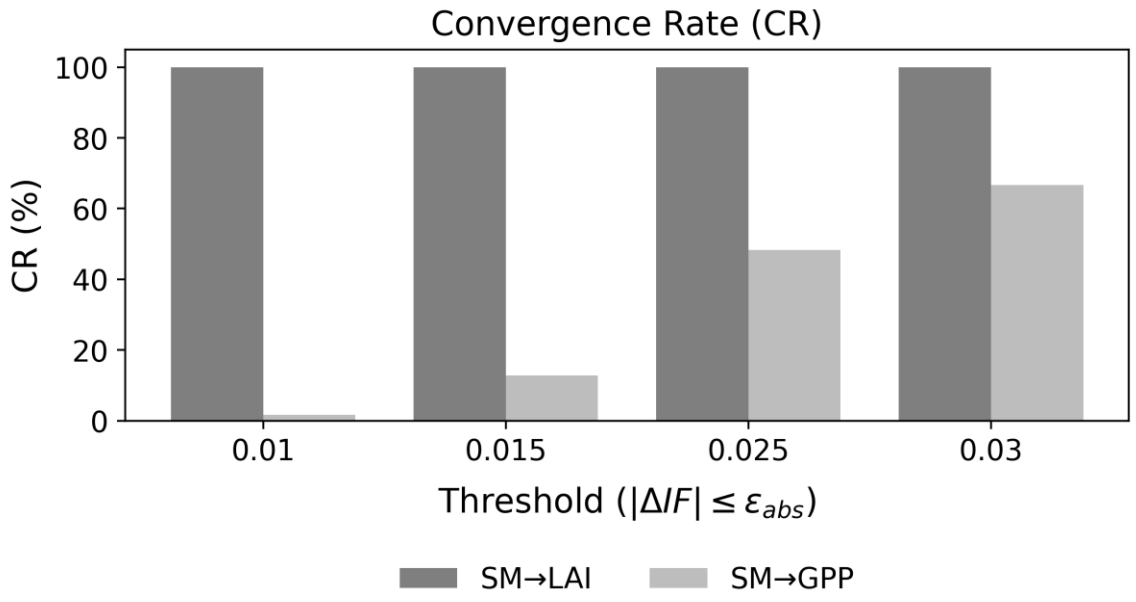


Supplementary Fig. S5. Sensitivity of the ANOVA decomposition to the number of quantile bins: terciles (top panel) and quintiles (bottom panel). ANOVA results showing the explained standard deviation (√AEV) of |ΔIF| for Southeast China. Columns represent the four pathways: SM→LAI, VPD→LAI, SM→GPP, and VPD→GPP. (a) Two-way ANOVA with factors SM and VPD. (b) Three-way ANOVA including SM, VPD, and T. (c) Three-way ANOVA including SM, VPD, and SSR. Bars show the explained SD for each factor or

factor interaction; error bars indicate the 95% confidence interval. (SM= Soil Moisture, VPD= Vapor Pressure Deficit, T= Temperature, SSR= Solar Surface Radiation). Note that the Y-axis range differs in each panel to better see the different values.

4. According to the authors, the definition of CR depends on a threshold ε_abs chosen as, for example, the 10th percentile of ΔIF across all couplings. This choice strongly affects CR yet is not sufficiently justified or stress-tested. Sensitivity analysis is needed because CR may change substantially under alternative thresholds.

Response: We thank the reviewer for raising this issue, which was also pointed out by the other reviewer. We agree that the CR definition depends directly on the selected absolute tolerance ε_abs , and this dependence was not sufficiently documented in the original manuscript. We have now clarified in Sect. 2.4.4 that the main analysis uses a fixed threshold of ε_abs =0.02, and we added a supplementary sensitivity analysis in which CR was recalculated for several alternative thresholds (ε_abs 0.01, 0.015, 0.025, and 0.03; Supplementary Fig. S6). The results show that, although the absolute CR value for SM→GPP varies substantially with the threshold, the main qualitative conclusion remains unchanged: SM→LAI remains consistently close to the fully conditioned estimate across all tested thresholds, whereas SM→GPP is much more sensitive to the tolerance and therefore more strongly mediated. We have revised the Methods and Discussion accordingly.



Supplementary Fig. S6. Convergence Rate (CR) to the absolute divergence threshold ε_{abs} for the Southeast China SM→LAI and SM→GPP couplings. CR was recalculated using ε_{abs} = 0.01,0.015,0.025, and 0.03. SM→LAI remains consistently near full convergence across all tested thresholds, whereas SM→GPP shows strong threshold dependence, although it remains substantially less convergent than SM→LAI.

5. The paper should report the employed Kalman filter parameter settings.

Response: We thank the reviewer for this important comment. We have now revised Section 2.1 to report the Kalman filter settings used in our implementation. The added paragraph reads:

In this study, the time-varying covariance matrix was estimated using a square-root Kalman filter within the Z24 framework. The Kalman gain was computed recursively within the filter from the evolving covariance estimates. The process-noise covariance matrix Q and measurement-noise covariance matrix RRR were estimated from the input series using a moving-window procedure. Specifically, the input vector was first smoothed using an unweighted moving average (UWMA) with a window length of 24 timesteps, and covariance estimation was then performed over a lookback window of 90 timesteps. The filter was initialized using the smoothed signal as the initial a posteriori state estimate and identity matrices as the initial covariance factors.

6. ANOVA is interesting, but readers need reassurance that its assumptions are not severely violated under quantile-conditioned, potentially heteroscedastic, non-independent data. A discussion on robustness is required.

Response: We thank the reviewer for this comment. We agree that ANOVA applied to quantile-conditioned |ΔIF| time series may not fully satisfy classical assumptions, particularly with respect to independence and homogeneity of variance. To address this, we have added a dedicated discussion on robustness in Appendix B. The added paragraph reads as follows:

While ANOVA provides a useful framework for partitioning variability in |ΔIF| across hydroclimatic regimes, it is important to note that its classical assumptions may not be fully satisfied in this context. The |ΔIF| values are derived from monthly time series and may therefore retain temporal dependence, meaning that observations are not strictly independent. In addition, the use of quantile-based regime definitions can lead to unequal variance across bins (heteroscedasticity), particularly in extreme or sparsely populated regime combinations. In this study, ANOVA is therefore interpreted primarily as a variance-partitioning and effect-size diagnostic, rather than as a strict inferential test under ideal assumptions. To assess robustness, we complement the analysis with alternative regime definitions (terciles and quintiles; Supplementary Fig. S5). The main qualitative conclusions, particularly the dominant role of higher-order interactions, remain consistent across these tests, although some weaker two-way effects show greater sensitivity to binning choice.

7. The new indices (MDI, MG, CP, and CR) are interesting, but they currently read more as descriptive summary statistics than as fully validated methodological advances. Their novelty would be stronger if the authors demonstrated formal properties, sensitivity (do results change if noise increases? time series length

changes? conditioning set slightly changes?), robustness, and interpretability (What does a high MDI physically mean?) across more than one showcased pathway.

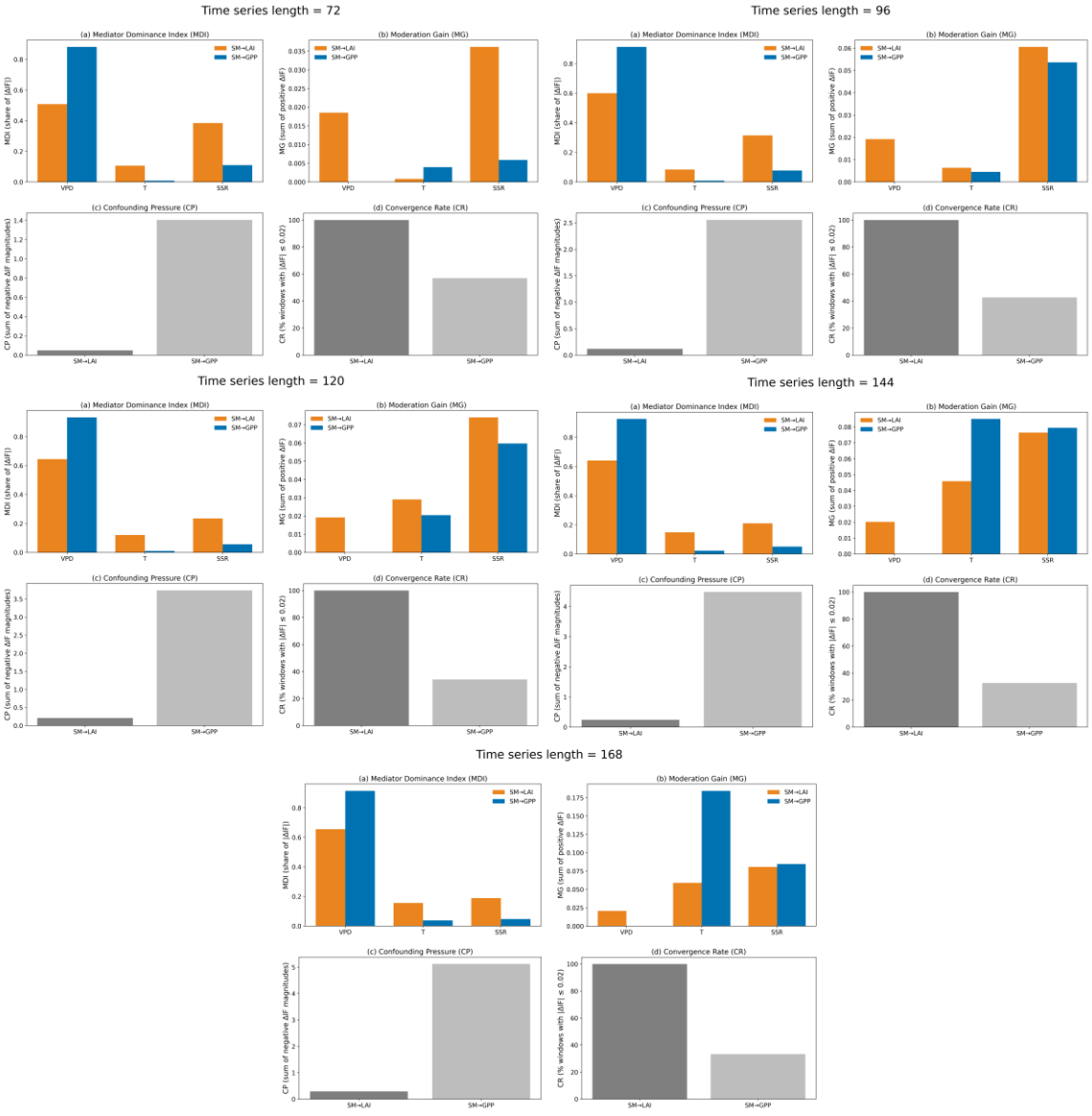
Response: We thank the reviewer for this thoughtful and constructive comment. In response, we have expanded the manuscript and Supplementary Information to include a more systematic empirical evaluation of the proposed indices (MDI, MG, CP, and CR) in terms of sensitivity, robustness, and interpretability.

Specifically, we assessed (i) sensitivity to perturbation by adding Gaussian noise (5–20% of the series standard deviation) to the stepwise IF-difference series and recomputing the indices across 200 realizations (Supplementary Table S1), and (ii) sensitivity to record length using truncated time series (72–168 months; Supplementary Fig. S8). In addition, we now clarify the physical interpretation of all indices in the revised Discussion. In particular, a high MDI identifies the conditioning variable that contributes most strongly to the observed |ΔIF| structure, while MG and CP quantify enhancement and suppression effects under conditioning, and CR reflects convergence between bivariate and conditioned estimates.

Together, these additions show that the proposed indices provide stable and interpretable diagnostic information across multiple pathways and under a range of sensitivity tests. A more formal theoretical characterization of their mathematical properties and a systematic assessment of sensitivity to alternative conditioning sets remain an important direction for future work.

Supplementary Table S1. Sensitivity of MDI, MG, CP, and CR to Gaussian noise of the stepwise ΔIF series. Values are reported as mean ± standard deviation across 200 realizations for noise levels corresponding to 5%, 10%, and 20% of the standard deviation of each series.

Pathway	Noise (%)	MDI_VPD	MDI_T	MDI_SSR	MG_VPD	MG_T	MG_SSR	CP	CR (%)
SM->GPP	5	0.9016 ± 0.0003	0.0466 ± 0.0002	0.0518 ± 0.0002	0.0 ± 0.0	0.2493 ± 0.0011	0.0849 ± 0.0008	5.4478 ± 0.0108	32.0389 ± 1.0525
SM->GPP	10	0.9008 ± 0.0006	0.0473 ± 0.0004	0.052 ± 0.0004	0.0 ± 0.0	0.2514 ± 0.0021	0.0855 ± 0.0017	5.4512 ± 0.0197	31.9111 ± 1.261
SM->GPP	20	0.8981 ± 0.0013	0.0493 ± 0.0008	0.0526 ± 0.0009	0.0001 ± 0.0004	0.2577 ± 0.0041	0.0874 ± 0.0034	5.4602 ± 0.0409	31.4583 ± 1.7341
SM->LAI	5	0.6574 ± 0.0013	0.1566 ± 0.0008	0.186 ± 0.0009	0.0308 ± 0.0006	0.0615 ± 0.0003	0.0817 ± 0.0004	0.2903 ± 0.0012	100.0 ± 0.0
SM->LAI	10	0.6567 ± 0.0023	0.1564 ± 0.0014	0.1868 ± 0.0017	0.0314 ± 0.0013	0.0617 ± 0.0005	0.0821 ± 0.0007	0.2914 ± 0.0023	100.0 ± 0.0
SM->LAI	20	0.6543 ± 0.0043	0.1566 ± 0.0027	0.1891 ± 0.0033	0.0337 ± 0.0025	0.0623 ± 0.001	0.0834 ± 0.0014	0.2958 ± 0.004	100.0 ± 0.0

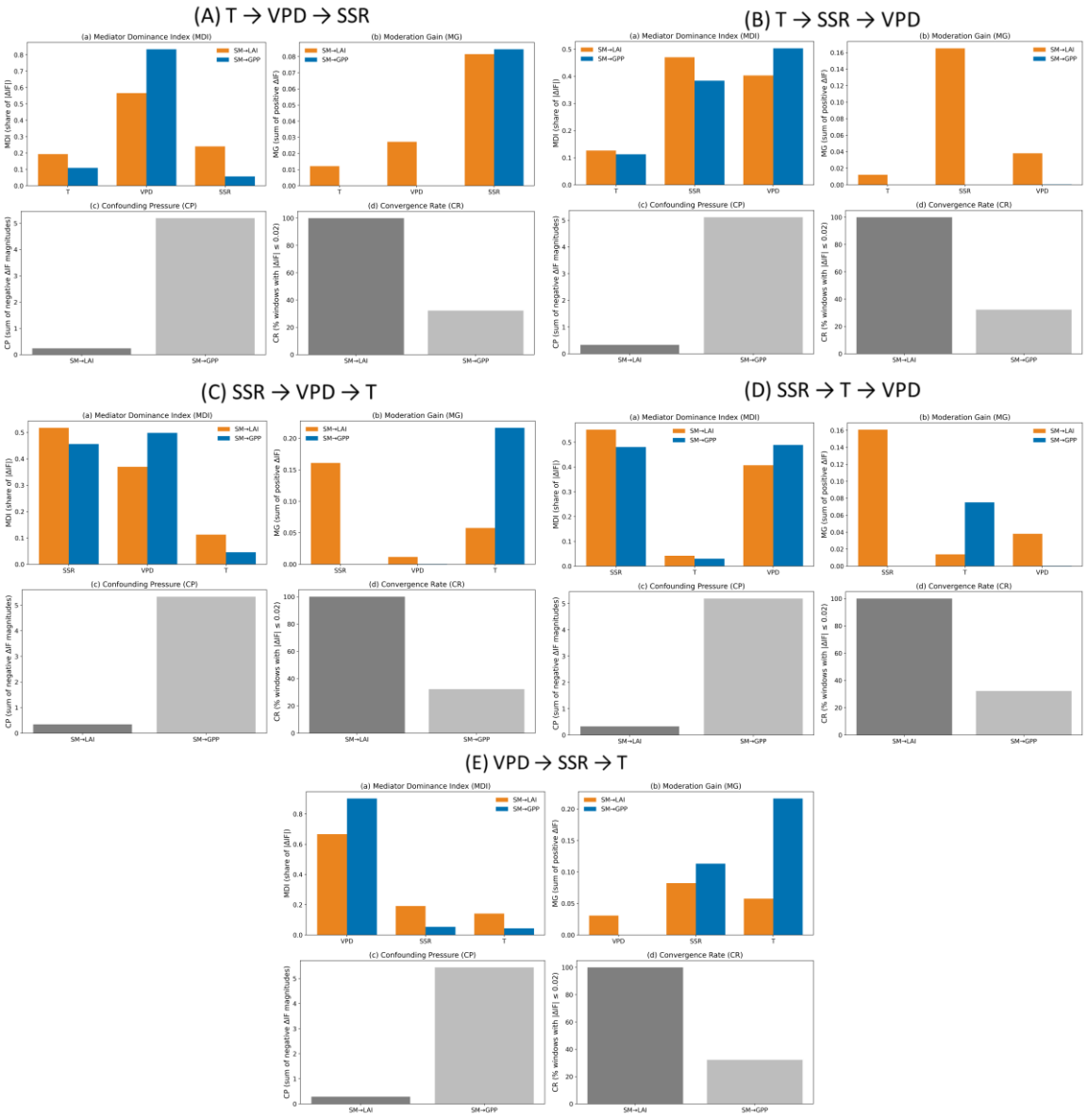


Supplementary Fig. S8. Sensitivity of MDI, MG, CP, and CR to time-series length. The indices were recomputed using records of 72, 96, 120, 144, and 168 months. While MG and CP show dependence on record length due to their cumulative definitions, the MDI and CR remain consistent across all lengths for both SM→LAI and SM→GPP.

Sensitivity to record length was evaluated by recomputing the indices using truncated time series (72–168 months). The dominant mediator identified by MDI remained consistent across all lengths, while MG and CP exhibited length dependence consistent with their cumulative definitions. CR remained persistently high for SM→LAI and substantially lower for SM→GPP across all lengths, indicating that pathway-level contrasts are stable with respect to sampling length.

8. The decomposition in Eqs. (7)–(9) is order-dependent, since conditioners are added progressively. The manuscript does not discuss whether MDI, MG, and CP change if the order of VPD, T, and SSR is permuted. This is a major issue because the estimated "contribution" of each conditioner may partly reflect the order rather than intrinsic causal relevance. The authors should test all permutations.

Response: We thank the reviewer for raising this issue. We agree that the decomposition in Eqs. (7)–(9) is order-dependent because the conditioners are introduced progressively. We therefore tested all permutations of the three conditioners (VPD, T, SSR) and added the results to Supplementary Fig. S7. These tests show that MDI and MG change with the order of introduction, confirming that they should be interpreted as order-dependent sequential attribution measures. By contrast, CR was unchanged across permutations, and the aggregated CP also remained unchanged in our application. We have revised the Methods to state this limitation explicitly and updated the Results/Discussion to qualify conclusions about mediator dominance using the permutation analysis. Please note that Reviewer 2 had a similar comment (see Major Comment 1).



Supplementary Fig. S7. Sensitivity of decomposition indices to conditioner ordering. MDI, MG, CP, and CR are computed for different permutations of the conditioning variables (VPD, T, SSR). Panels (A–E) represent alternative ordering sequences: (A) T → VPD → SSR, (B) T → SSR → VPD, (C) SSR → VPD → T, (D) SSR → T → VPD, and (E) VPD → SSR → T. MDI and MG vary with permutation order, indicating order dependence, whereas CR remains invariant and the aggregated CP remains unchanged in this application. Results are shown for both SM–LAI and SM–GPP pathways.

9. Careful language editing is required throughout the manuscript. Examples include awkward phrasing such as “This begs a spatio–temporal quality…,” “more direct SM→LAI,” “VPD_driver-based,”.

Response: We thank the reviewer for noting these language issues. The manuscript has been carefully edited throughout to improve clarity, grammar, and scientific phrasing.