

Dear Editor,

We sincerely thank you and the three reviewers for the careful evaluation of our manuscript and for the constructive comments and suggestions. We have revised the manuscript accordingly and provided detailed point-by-point responses to all comments. For convenience, the responses to all three reviewers have been compiled into a single PDF. Throughout the response document, reviewer comments are shown in black, our responses in blue, and the corresponding revisions in the manuscript are highlighted in yellow.

Sincerely,

Referee #3

This manuscript applies the multivariate, time-varying LKIF framework to compare bivariate and multivariate information-flow estimates via ΔIF and related diagnostics (MDI, MG, CP, CR), focusing on SM/VPD $\rightarrow$ LAI/GPP couplings in Southeast China and the Amazon Basin. The topic is relevant and the paper is generally readable. The revision adds useful material on Kalman settings, CR thresholds, quantile sensitivity, conditioner-order permutations, and an expanded Discussion.

However, when the original manuscript is read together with the point-by-point response, several fundamental issues remain inadequately addressed. Responses often rely on wording changes, supplementary sensitivity plots, or deferral to future work, rather than targeted analyses. Although Liang–Kleeman information flow (LKIF) is now used widely across atmospheric, oceanic, financial, and other applications, the central claim of this paper—that the bivariate–multivariate IF difference can characterize whether a physically meaningful mechanism is present between driver and response—is not yet sufficiently supported in terms of causal identification and synthetic validation. Major revision is required before the main conclusions can be considered robust:

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.

1. Title overstates scope; representativeness concern not substantively answered

The title “Benchmarking Conditioners in Liang–Kleeman Information Flow” suggests a systematic benchmark, yet the study is a limited application to two small humid-forest domains and a handful of variables/pathways. There is no comprehensive or global analysis, and Appendix A does not constitute a conditioner benchmark in the strict sense. It is recommended that using a more modest title (e.g., ... A Case Study of Land–Atmosphere Interactions in Humid Forests, or ... A Simple Application to Land–Atmosphere Interactions); tone down global claims; state explicitly that real-data inferences are restricted to humid forests.

Besides, reviewer 2 raised this explicitly (Major Comment 5). In review.pdf, the authors mainly reworded the Abstract/Conclusions and flagged broader biomes as future work. No additional regions or biome contrasts were added. The revised Abstract sentence—“even for the same variables ... differently at different locations”—emphasizes spatial heterogeneity but does not answer whether conclusions from two humid forests generalize elsewhere. This reads as deflecting the representativeness criticism rather than addressing it.

Response: We thank the reviewer for the comment. We agree that the ecological scope of the real data application should be made explicit in the title and throughout the manuscript. We have therefore revised the title to:

Benchmarking Conditioners in Liang–Kleeman Information Flow: Application to Land–Atmosphere Interactions in Humid Forests

We further clarify, however, that the term “**benchmarking**” is intended to refer to the **methodological evaluation of conditioning effects within the LKIF framework**, rather than to a comprehensive benchmark across global biomes. The study compares bivariate and multivariate time-varying LKIF estimates, evaluates their divergence, and introduces diagnostic measures to quantify how conditioning variables modify the inferred information flow. In this sense, the benchmarking concerns the behavior of the methodology under different conditioning configurations. We do not expect the fundamental lessons learned here to change significantly upon a different application for which we have already verified.

Additionally, the real-world analysis serves as an application of this methodological framework and is explicitly restricted to two humid-forest regions: Southeast China and the Amazon Basin. We agree that these two regions do not provide a basis for global biome-level generalization. However, the purpose of including them is not to benchmark the importance of conditioners across ecosystems, but to demonstrate how the proposed diagnostics behave in two geographically distinct humid-forest systems to verify our hypothesis: conditioner importance for causality structure will vary across two or more locations, even when given the same variables across the same time-frame. Additionally, even for the same location, their importance will vary across time for the same variables. This is a relevant lesson for all causal frameworks since most of them assume stationarity of causal structure, including the classic form of the LKIF.

We would also like to clarify that the absence of additional biomes was not an omission in our response to the previous review. Reviewer 2 recommended either adding additional representative biome regions **or, if that was not feasible, explicitly identifying the biome context and limiting the conclusions accordingly**. We followed the second option because adding drylands, boreal forests, and other contrasting ecosystems would require a substantially broader comparative design and would extend beyond the scope of the present study. And more importantly, this is already part of an ongoing study.

Accordingly, the revised manuscript:

1. identifies both study domains as humid forests;
2. restricts the empirical interpretation to these two regions;
3. avoids claiming that the observed conditioner importance is representative of all biomes; and
4. acknowledges that other ecosystems may exhibit different conditioner effects and causal structures.

We agree that the previous Abstract sentence concerning differences between locations addressed spatial heterogeneity but did not directly address representativeness. We have therefore revised the Abstract to identify the two application regions explicitly as **Southeast China and the Amazon Basin**. In addition, we have added a clear statement to Section 3.1 that the results presented in the study are specific to these two humid forest regions and should not be generalized to other regions or biome types. The revised text reads as,

The results presented in this study are based on these two humid-forest regions; however, they hold broader lessons that are relevant for stationary assumptions and conditioner inclusions for the LKIF. In fact, a recent application of these findings to improving predictability using machine learning also points to the importance of the study findings (Hagan et al., 2026; <https://openreview.net/forum?id=Mt5MIMVowi#discussion>).

2. For Southeast China $SM \rightarrow GPP$ (Fig. 1c), bivariate IF is mostly insignificant except around 2011, whereas $SM \rightarrow GPP|VPD,T,SSR$ is significant in multiple periods. Conditioning therefore changes both magnitude and significance. Appendix A instead emphasizes directional fidelity and near-zero false links, with bivariate and multivariate curves nearly overlapping for $x_1 \rightarrow x_2$. It does not reproduce significance transitions or large conditioning effects. This is not a strict logical

contradiction, but a major evidential gap: synthetic validation does not support the paper's central real-data message.

Response: We thank the reviewer for this comment. We agree that Appendix A does not capture the full range of conditioning effects highlighted in the observational application, particularly the conditioning-induced changes in statistical significance.

We would like to clarify, however, that the negligible adjustment between the bivariate and multivariate estimates depicted in the appendix is specific to the $x_1 \rightarrow x_2$ pathway. In the same synthetic experiment, conditioning on x_1 changes the magnitude of the estimated IF for $x_3 \rightarrow x_2$, while the inferred causal direction remains unchanged. Thus, the synthetic experiment shows that the response to conditioning differs across the pathways considered.

We nevertheless agree that the present synthetic configuration does not exhibit the type of conditioning-induced transitions in statistical significance highlighted by the reviewer. Appendix A should therefore not be interpreted as reproducing the full range of conditioning behavior found in the observational application. Its purpose is more limited and focuses on examining LKIF behavior under the specific controlled synthetic configuration considered.

We have clarified this point in Appendix A as:

Importantly, the bivariate and multivariate IF curves for $x_1 \rightarrow x_2$ remain nearly indistinguishable throughout the entire γ range, indicating that conditioning on x_3 provides only a negligible adjustment for this pathway. In contrast, for $x_3 \rightarrow x_2$, conditioning on x_1 changes the magnitude of the estimated IF, while the inferred causal direction remains unchanged (Fig. A1c). These results show that the response to conditioning differs between the pathways considered in the synthetic experiment. The present synthetic configuration, however, does not exhibit conditioning-induced transitions in statistical significance between the bivariate and multivariate estimates.

3. Appendix A does not systematically test chain, observable fork, collider, direct+indirect, or joint multivariate conditioning as in the main text. For example, if X_5 is observed but X_4 is an unobserved mediator on $X_2 \rightarrow X_4 \rightarrow X_5$, can LKIF in subsystem (X_1, X_2, X_3, X_5) avoid misattributing X_4 's role? If $X_1 \rightarrow X_2 \rightarrow X_4$ and $X_1 \rightarrow X_3 \rightarrow X_4$ contribute with opposite signs and similar magnitudes, bivariate $T(X_1 \rightarrow X_4)$ may cancel, and $T(X_1 \rightarrow X_4 | X_2, X_3)$ may remain similarly small. Then $|\Delta IF|$ is

small and CR is high even though X_2 and X_3 are essential—contradicting the claim that small $|\Delta IF|$ implies bivariate sufficiency.

Response: Thank you. We agree that Appendix A does not systematically test all possible causal structures, including chains, observable forks, colliders, hidden mediators, joint multivariate conditioning, and parallel pathways with compensating effects. The purpose of Appendix A was to examine LKIF behavior under the specific controlled hidden-confounding configuration considered there, rather than to provide an exhaustive assessment of causal graph recovery.

We also agree with the reviewer’s point concerning the interpretation of small $|\Delta IF|$. In this study, a small $|\Delta IF|$ indicates that the bivariate form is very similar to the multivariate one. However, such similarity does not establish causal sufficiency or imply that individual conditioning pathways are unimportant. As the reviewer correctly notes, different pathways may contribute with opposite signs and similar magnitudes, so that their effects compensate and yield a small overall difference between the bivariate and multivariate estimates despite substantial contributions from individual pathways. Conversely, a large $|\Delta IF|$ suggests that the estimated causal influence is substantially modified when additional variables are introduced.

The reviewer’s example involving an unobserved mediator raises a related limitation. Appendix A does not test whether the contribution of an unobserved mediator can be uniquely separated when that mediator is absent from the observed subsystem. We have therefore clarified both the scope of the synthetic experiment and the interpretation of small $|\Delta IF|$.

We have added the following paragraph to Appendix A:

The present synthetic experiment is intended as a controlled assessment of LKIF under the specific hidden-confounding structure considered here and does not constitute a systematic benchmark across all possible causal structures. In particular, configurations involving hidden mediators, colliders, observable forks, joint conditioning structures, or parallel pathways with compensating effects are not systematically evaluated. This distinction is important for interpreting $|\Delta IF|$. A small $|\Delta IF|$ indicates that the bivariate and multivariate estimates are very similar for the specified conditioning set but does not establish causal sufficiency. For example, different pathways may make individually important contributions with opposing signs and similar magnitudes, resulting

in a small overall difference. Similarly, when a relevant mediator is unobserved, the present experiment does not establish that its contribution can be uniquely separated within the observed subsystem. A systematic assessment of LKIF under these different causal structures represents an important direction for future work.

Section 2.2 has been revised to:

A small $|\Delta IF|$ indicates that the bivariate form is very similar to the multivariate one. In such cases, the inclusion of additional variables does not notably alter the overall estimated causal influence. However, this similarity does not establish causal sufficiency or imply that individual conditioners are unimportant, because opposing pathway contributions may compensate and yield a small overall difference.

We have further added this limitation to the Discussion and revised the corresponding statements in the Abstract, Results, and Conclusion to avoid interpreting small $|\Delta IF|$ or high CR as evidence of directness or causal sufficiency. These cases are now described in terms of similarity between the bivariate and multivariate estimates or weak sensitivity to the selected conditioning variables.

4. Land–atmosphere memory often spans weeks to months. The study uses monthly anomalies and contemporaneous LKIF via Kalman covariances, with no explicit lag structure.

Response: Thank you for this comment. We agree that land–atmosphere interactions can exhibit memory over timescales ranging from weeks to months, and that an explicit lagged analysis could provide additional information on delayed driver–response relationships.

The present study, however, was designed to investigate how conditioning variables modify time-varying LKIF estimates, rather than to identify characteristic response lags or reconstruct the full temporal memory structure of land–atmosphere interactions. The monthly resolution is expected to provide an integrated representation at the temporal scale considered in this study, including variability arising from processes operating at shorter timescales, while allowing us to focus specifically on the role of conditioners.

We nevertheless agree that lagged relationships were not examined in the present analysis. We have therefore acknowledged this aspect in the Conclusion as follows:

The present analysis does not evaluate lag-specific driver–response relationships. Given that land–atmosphere interactions may exhibit memory over weeks to months, future work should investigate how the inferred information flows and the influence of conditioning variables vary across different temporal lags.

5. Fig. 2: High/low ΔIF periods are illustrated with one year each for causal graphs, without demonstrating representativeness or a formal selection algorithm beyond post-hoc labeling. Why not composite graphs from all high- $|\Delta IF|$ or low- $|\Delta IF|$ years? Single years may reflect outliers, extremes, or estimator artifacts rather than typical architecture.

Response: We thank the reviewer for raising this point. The time points shown in Fig. 2 were intentionally selected as illustrative examples of contrasting cases of large and small $|\Delta IF|$, rather than as composite or statistically representative causal graphs for all high- and low- $|\Delta IF|$ periods. Our purpose was to examine how the complete five-variable causal structure behaves when the bivariate and multivariate estimates for the SM/VPD→LAI/GPP couplings show particularly strong divergence or convergence. This is consistent with the methodological focus of the manuscript, which is to investigate when and why bivariate and multivariate estimates differ, rather than to establish a typical mechanistic architecture for every environmental regime.

The selected time points were also considered in relation to the documented hydroclimatic conditions during the corresponding periods. In Southeast China, the selected 2011 period coincided with the persistent and severe winter–spring drought documented by Sun and Yang (2012, doi:10.1029/2012JD017500). In the Amazon, the selected 2014 time point occurs during a period in which changes in forest recovery dynamics have been reported. Van Passel et al. (2024, doi:10.1073/pnas.2316924121) reported an increasing critical-slowness signal from approximately 2013–2014 onward and associated the decline in forest recovery capacity with increasing drought occurrence. Major Amazon droughts and their associated impacts on forest functioning have also been documented by Lewis et al. (2011, doi:10.1126/science.1200807) and Feldpausch et al. (2016, doi:10.1002/2015GB005133). Therefore, the selected graphs were intended to illustrate how pronounced divergence or convergence is reflected in the broader five-variable causal structure under environmentally contrasting conditions and, in the high-divergence cases, during periods of hydroclimatic stress. They are not interpreted as representative of every

high- or low- $|\Delta IF|$ period, and the broader conclusions are not derived solely from these individual years.

We did not construct composite graphs because the presence, direction, and strength of individual causal links vary over time, which is the main hypothesis of the study. Combining all high- or low- $|\Delta IF|$ periods would require averaging potentially different or opposing links or applying an additional threshold to determine which links should be retained. Such a composite could overcomplicate the time-varying causal structure that the methodological framework is intended to reveal.

The following paragraph has been added to Section 3.2 to clarify this point:

The selected time points were chosen to illustrate contrasting cases of pronounced large and small $|\Delta IF|$. For the large- $|\Delta IF|$ cases, the selection also considered the documented hydroclimatic and ecological conditions during the corresponding periods. In Southeast China, the selected 2011 period coincided with the persistent and severe winter–spring drought reported by Sun and Yang (2012). In the Amazon, the selected 2014 time point occurred during a period in which changes in forest recovery dynamics have been reported, with Van Passel et al. (2024) identifying an increasing critical-slowness trend from around 2014 onward and relating slowing-down patterns to increased drought occurrence. The selected graphs are intended as illustrative examples of contrasting ΔIF conditions rather than statistically representative composites of all high- and low- $|\Delta IF|$ periods. Because the presence, direction, and strength of individual causal links vary through time, combining multiple periods into a composite graph could average or obscure distinct and potentially opposing causal links.

6. Mapping $+IF/-IF$ to vulnerability/resilience is not part of the strict LKIF definition. If intended as an ecological extension, it requires independent justification and citations; otherwise it belongs in Discussion as heuristic interpretation—or a separate paper.

Response: We thank the reviewer for this helpful comment. We agree that directly mapping positive and negative LKIF to ecological vulnerability and resilience extends beyond the strict definition of the Liang–Kleeman information flow and should not be presented as an intrinsic property of the method.

Within the LKIF framework, IF is a signed rate of directed information transfer. A statistically significant nonzero $IF_{X \rightarrow Y}$ indicates a directed influence of X on Y, while its magnitude characterizes the strength of that inferred influence. The sign has a more specific information-theoretic meaning: positive IF indicates that variability in the source contributes to increasing the variability of the target, whereas negative IF indicates a variability-reducing contribution. Importantly, the sign does not indicate whether the mean value of the target increases or decreases. This interpretation is consistent with Docquier et al. (2024), who explicitly distinguish the magnitude of LKIF as the strength of the causal influence from the sign, which reflects enhancement or reduction of target variability.

Our original use of the terms “vulnerability” and “resilience” was motivated by the broader ecological literature linking information-theoretic concepts with ecosystem stability. In particular, Rutledge et al. (1976, doi: 10.1016/0022-5193(76)90007-2) provide conceptual precedent for relating information-theoretic properties of ecological systems to their response to perturbations. We agree, however, that this literature does not establish a direct equivalence between the sign of LKIF and ecological vulnerability or resilience.

We have therefore revised the manuscript to clearly separate these two levels of interpretation. The general methodological description now uses only the LKIF-supported terminology of variability-enhancing and variability-reducing influence, and we have removed the general mapping of +IF to vulnerability and –IF to resilience, as well as the associated references to “extreme states” and “equilibrium” as formal interpretations of IF. Ecological terms such as vulnerability or resilience are retained only where they are independently supported by the physical and ecological context of the specific result and are presented as contextual interpretations rather than as properties of the LKIF sign itself.

The interpretation of Information Flow in Section 2.1 has been revised as follows:

We interpret LKIF as a signed rate of directed information transfer from a source variable to a target variable. A statistically significant nonzero LKIF indicates a directed causal influence of the source on the target within the LKIF framework, while the absolute magnitude of LKIF characterizes the strength of the directed influence. When LKIF is positive (+IF), this suggests that

variability in the source increases the variability of the target. In contrast, a negative LKIF ($-IF$) value indicates that variability in the source decreases the variability of the target (Liang, 2014; Docquier et al., 2024).

7. Fig 3: In cells such as SM(Q1)–VPD(Q1), $|r_{IF}|$ appears blank/white while $|\Delta IF|$ is non-zero. The caption explains cross-hatching for “no data” but not why correlation is unavailable while $|\Delta IF|$ is shown (likely small n , zero variance, or inconsistent computation rules). This undermines Section 3.3.

Response: We thank the reviewer for identifying this ambiguity. The reviewer is correct that the blank/white $|r_{IF}|$ triangle in the SM(Q1)–VPD(Q1) regime is due to the small sample size. Specifically, only two paired observations ($n=2$) were available for this regime during the study period. Although $|\Delta IF|$ can be calculated from these paired observations, a Pearson correlation based on only two points is not statistically informative because it necessarily yields $|r|=1$. We therefore did not report $|r_{IF}|$ for this regime, while retaining the corresponding $|\Delta IF|$ estimate.

We have revised the caption of Figure 3 and Supplementary Figure S3 to explain this convention. We have also ensured that these cells are not used to support the correlation-based interpretations in Section 3.3.

Addition to the captions of Figure 3 and Supplementary Figure S4:

Blank upper triangles, such as that for SM(Q1)–VPD(Q1), indicate quartile combinations for which only two paired observations ($n=2$) were available; in these cases, $|r_{IF}|$ was not reported because Pearson correlation is not statistically informative with $n=2$, while the corresponding $|\Delta IF|$ value was retained.

8. Because it is impossible in practice to include the full Earth-system state vector in LKIF computation, and because currently implementable IF algorithms—whatever the generality of the underlying theory—operate under linear approximations in applied settings, the authors should use numerical simulation more systematically to validate the causal pathways inferred from observations. Without such validation, it is difficult to distinguish physically meaningful conditioning effects from artifacts of omission, linearization, or sparse sampling.

Response: Thank you. We agree that, in complex Earth-system applications, it is impossible to include the complete state vector in an LKIF analysis, and that practical implementations necessarily involve assumptions related to variable selection, model approximation, and data availability. In fact, this is a limitation for every existing causal inference or causal discovery framework.

The objective of this work is not to reconstruct all possible causal pathways or to provide a complete causal representation of land–atmosphere systems. Instead, we focus specifically on understanding the role of conditioners (mediators and confounders) in modifying LKIF estimates. We investigate when and why bivariate and multivariate LKIF estimates diverge, how individual conditioning variables contribute to this divergence, and under which environmental conditions conditioning substantially alters the inferred causal strength.

In this context, the selected conditioners (SM, VPD, temperature, and surface solar radiation) are not intended to represent the complete Earth-system state vector, but rather key physically motivated variables known to regulate land–atmosphere interactions. By comparing bivariate and conditioned LKIF estimates, this study quantifies the sensitivity of inferred causal relationships to the inclusion of these relevant environmental factors.

We acknowledge that remaining unobserved processes and nonlinear interactions may influence the estimated information flows. Therefore, the synthetic experiment in Appendix A was designed to evaluate the fundamental behavior of LKIF under controlled hidden-confounding conditions, rather than to validate every possible Earth-system configuration. A systematic numerical validation framework incorporating the complete range of Earth-system processes, nonlinear interactions, and all potential missing variables would require the development of a separate process-based causal modeling framework. Such an analysis would substantially broaden the scope of the current study and shift the focus away from our primary objective, which is to quantify the role of conditioners in modifying LKIF-based causal estimates.

Although such a comprehensive modelling evaluation is beyond the scope of the present study, we agree with the reviewer that, in parallel with the current observation-based analysis, there is a need for detailed process-based modelling analyses to further evaluate the inferred causal pathways and

conditioning effects. Such numerical experiments would allow relevant processes and causal structures to be examined under controlled conditions and would help assess the influence of omitted processes, nonlinear interactions, and sampling limitations on the inferred information flows. We therefore consider this an important complementary direction for future research.

The Conclusion has therefore been revised to emphasize this point as follows:

Nevertheless, the inferred pathways remain conditional on the variables included, the linear formulation used here, and the sampling characteristics of the observations. In parallel with observation-based causal analysis, future work should therefore use targeted process-based numerical experiments to evaluate the identified pathways and conditioning effects under controlled changes in omitted processes, nonlinear interactions, and sampling characteristics. Such complementary modelling analyses would help assess the physical robustness of the inferred conditioning effects. Future work should also explore nonlinear extensions of the IF framework (Pires et al., 2024; Vannitsem et al., 2025) to better capture complex feedbacks that arise during extreme events.

Referee #4

Review of manuscript MS No.: egusphere-2026-267 Title: Benchmarking Conditioners in Liang–Kleeman Information Flow: Application to Land–Atmosphere Interactions
Author(s): Fareeha Siddique, Daniel Fiifi Tawia Hagan, Guojie Wang, David Docquier, and Stéphane Vannitsem

The authors present an interesting study of temporal evolution of the Liang–Kleeman Information Flow (LKIF) in its bivariate and multivariate forms in land–atmosphere interactions. I was asked for an opinion after the first revision of the manuscript and found that the first two reviews motivated the authors to make a great improvement of the manuscript. I think, however, that there are still some questions to be answered.

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.

1. The application of the Kalman filter allows for computation of the time-resolved LKIF. Later the authors discuss how the LKIF and its differences depend on values of involved variables, using combinations of their quartiles. Considering the LKIF is computed from co/variances, it would be useful to plot means and especially variances of the involved variables in the same time-resolution as LKIFs in order to see whether the time-evolution of LKIFs can or cannot be explained by the changes in variances.

Response: Thank you for this comment. The fundamental theoretical question of whether the time evolution of LKIF can or cannot be explained by the changes in variances (and covariances) has already been explicitly addressed in prior foundational literature (Liang, 2014; doi: 10.1103/PhysRevE.90.052150, Liang, 2021; doi: 10.3390/e23060679 and Zhou et al., 2024; doi: 10.1175/JCLI-D-23-0207.1).

These studies formally demonstrate that LKIF captures directional information transfer and rate-of-change dynamics rather than simply being explained by the changes in variances alone.

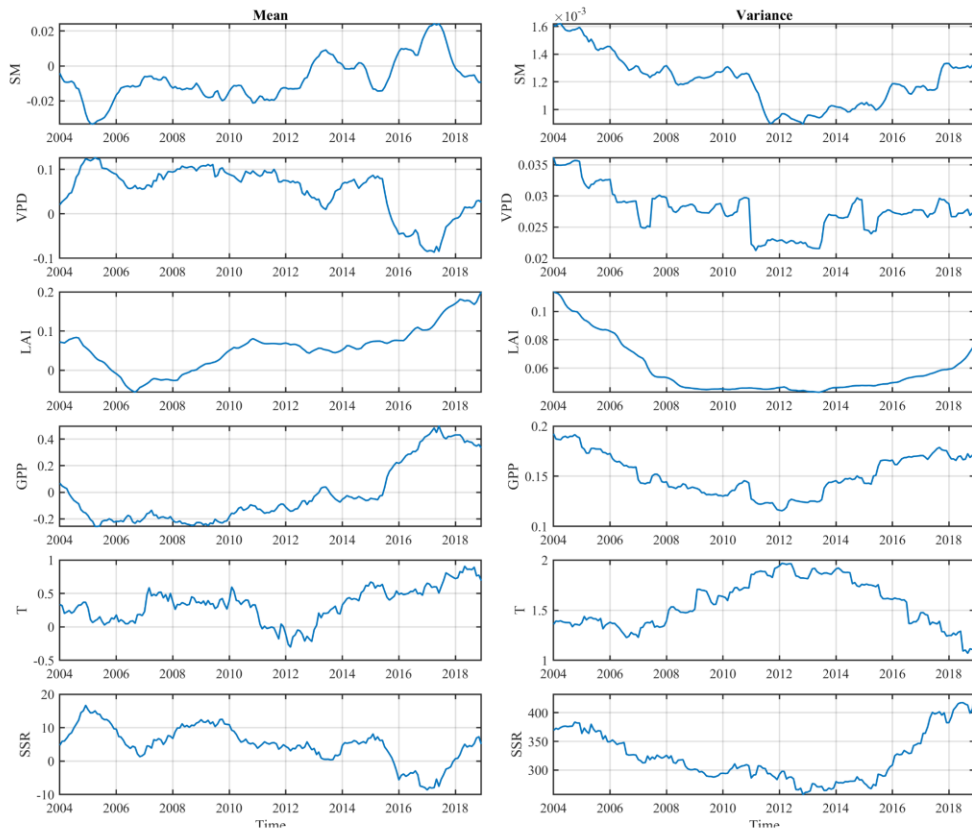
To directly address the reviewer's concern in the present analysis, we additionally calculated the time-resolved means and variances of all variables involved in the LKIF systems, including SM, VPD, LAI, GPP, T, and SSR, for both Southeast China and the Amazon Basin. The calculations were performed using the same anomaly time series and temporal settings used for the time-resolved LKIF estimation. The resulting time series have been added to the Supplementary Material (Supplementary Fig. S2).

Comparison with the LKIF time series in Fig. 1 shows that some periods exhibit co-variation between LKIF and the variance of individual variables, but this relationship is not consistent across time, pathways, or regions. For example, in the Amazon Basin, the increase in VPD variance coincides with part of the strengthening of the VPD-related information flows, but this correspondence does not persist later in the record, when VPD variance remains relatively high while the corresponding IF estimates decrease or change substantially. The SM-driven pathways likewise show periods in which changes in IF differ from the temporal evolution of SM and vegetation variances. These comparisons indicate that changes in individual variances may contribute to the temporal variability of LKIF but do not alone explain its observed evolution.

The Results section has been revised to include this comparison as follows:

To assess whether the temporal evolution of LKIF could be attributed to changes in the underlying variable variances, we additionally evaluated the time-resolved means and variances of all variables included in the analysis (SM, VPD, LAI, GPP, T, and SSR; Supplementary Figs. S2). Although some periods show co-variation between LKIF and individual-variable variance, no consistent correspondence is observed across pathways, time, and regions. This indicates that changes in individual variances alone do not explain the temporal evolution of LKIF.

(a) Southeast China



(b) Amazon Basin

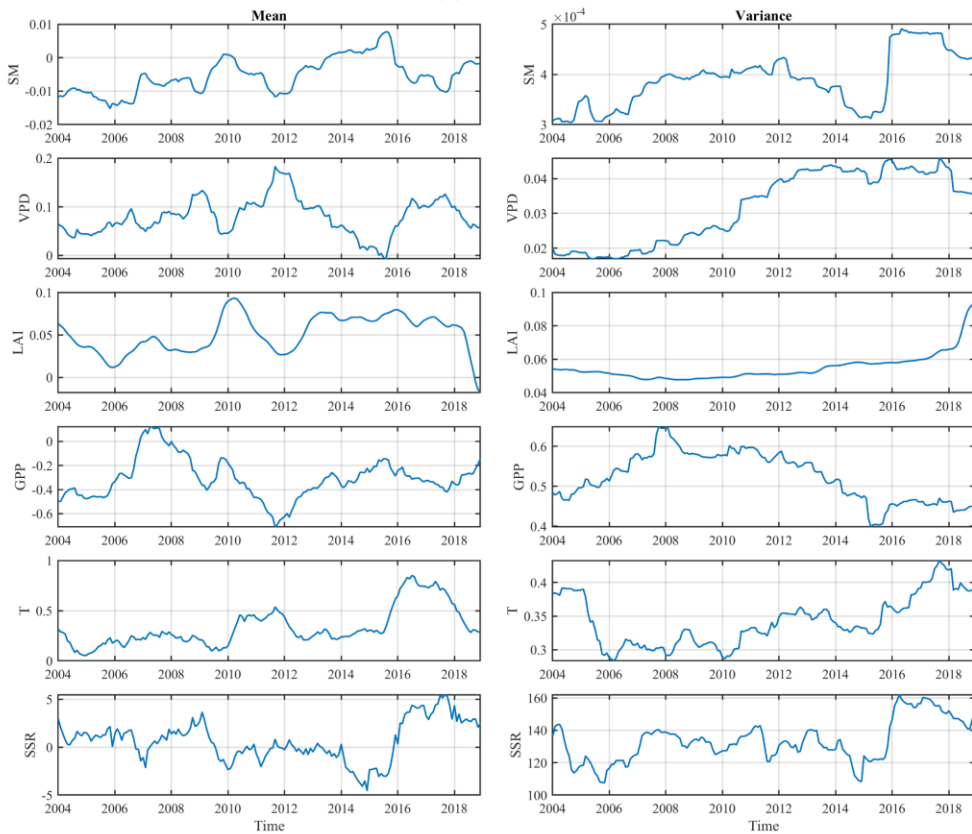


Fig. S2. Temporal evolution of the means (left column) and variances (right column) of SM, VPD, LAI, GPP, T, and SSR for (a) Southeast China and (b) the Amazon Basin during 2004–2018. The statistics were calculated from the same anomaly time series and using the same temporal settings as in the LKIF analysis (Materials and Methods). The units of the mean time series are $\text{m}^3 \text{m}^{-3}$ for SM, kPa for VPD, $\text{m}^2 \text{m}^{-2}$ for LAI, $\text{g C m}^{-2} \text{d}^{-1}$ for GPP, $^{\circ}\text{C}$ for T, and W m^{-2} for SSR. The variances are expressed in the corresponding squared units.

2. Typically, the causality measures decrease after conditioning (bivariate - multivariate > 0) as we can see also in the presented numerical example A3, while in some real data cases in Fig. 1 it is reversed, e.g., $\text{SM} \rightarrow \text{LAI} < \text{SM} \rightarrow \text{LAI} | \text{VPD}$. How this can be explained?

Response: Thank you. We agree that the original synthetic example may give the impression that conditioning should generally reduce the estimated IF. However, the reduction observed in Appendix A reflects the particular pathway configuration used in that experiment and is not a general constraint of conditional LKIF.

To examine this directly, we performed an additional controlled experiment using the same VAR (1) system but reversed the sign of the $x_1 \rightarrow x_2$ coefficient from $C_{12}=0.45$ to $C_{12}=-0.45$, while retaining all other parameters. This produces opposing direct $x_3 \rightarrow x_2$ and mediated $x_3 \rightarrow x_1 \rightarrow x_2$ pathways. Under this configuration, the conditioned information flow $T_{3 \rightarrow 2|1}$ is consistently larger than the bivariate $T_{3 \rightarrow 2}$ across the full range of hidden-confounder strengths. Correspondingly, ΔIF is negative at all tested values of γ , with the paired bootstrap 95% confidence intervals excluding zero.

This additional experiment demonstrates that conditioning does not necessarily lead to a reduction in LKIF. Depending on the signs and structure of the interacting pathways, accounting for an additional variable can either decrease or enhance the estimated driver–target information flow. The reverse ordering observed in Fig. 1 is therefore permissible within the LKIF framework and is consistent with the presence of opposing or masking contributions among the environmental variables. We note, however, that the observational analysis does not uniquely establish that this specific pathway configuration is responsible for each reversal seen in Fig. 1.

We have made this clear in Appendix A and added the corresponding controlled experiment as Supplementary Fig. S12 as follows:

The reduction of IF after conditioning in Fig. A1(c) reflects the particular pathway configuration prescribed in this synthetic experiment and is not a general requirement of conditional LKIF. To examine the reverse case, we repeated the experiment with the sign of the $x_1 \rightarrow x_2$ coefficient reversed from $C_{12}=0.45$ to $C_{12}=-0.45$, while retaining all other model parameters. This makes the direct $x_3 \rightarrow x_2$ pathway and the indirect pathway through x_1 act in opposite directions. Under this configuration, the conditioned IF $T_{3 \rightarrow 2|1}$ consistently exceeds the corresponding bivariate $T_{3 \rightarrow 2}$ estimate, yielding negative ΔIF across the full range of confounder strengths (Supplementary Fig. S12). Thus, conditioning can either reduce or enhance the estimated IF depending on the structure and signs of the interacting pathways.

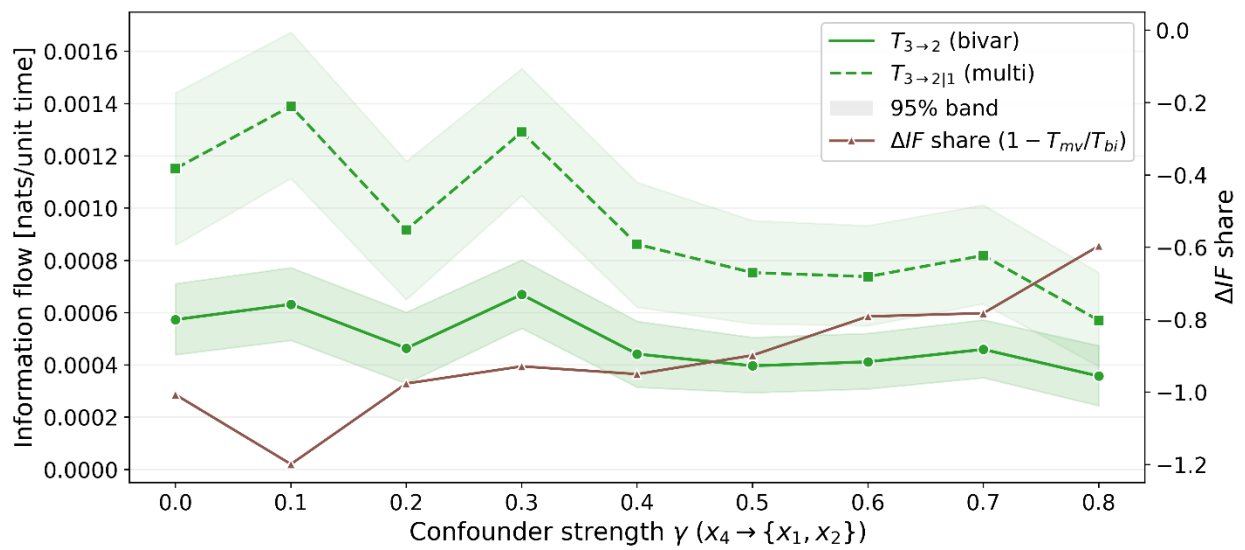


Fig. S12. Synthetic LKIF experiment illustrating enhancement of information flow after conditioning. The VAR(1) configuration is identical to that used in Appendix A except that the $x_1 \rightarrow x_2$ coefficient is reversed from $C_{12}=0.45$ to $C_{12}=-0.45$, causing the direct $x_3 \rightarrow x_2$ pathway and the indirect pathway through x_1 to act in opposite directions. The conditioned $T_{3 \rightarrow 2|1}$ exceeds the corresponding bivariate $T_{3 \rightarrow 2}$ throughout the tested confounder-strength range. The brown line shows the normalized difference (right y-axis). The shaded band shows the 95% confidence intervals.

3. Usually in causality analysis both directions are analysed, here only the causal directions from meteo/hydro to bio- variables are presented. Was this an intension? The only exception are the causal graphs in Fig 2, where also arrows e.g. from LAI to T can be seen. Were the latter obtained from five-variate estimates?

Response: Thank you. The focus on causal directions from SM and VPD to LAI and GPP was intentional, as these represent terrestrial water supply and water demand, which can constrain vegetation structure and function. These specific directions are of key importance and are the subject of debate within the field of land–atmosphere interactions, and therefore represent a key scientific interest to the community. Nonetheless, in this study, the main objective of this manuscript is methodological science, with a focus on examining when and why bivariate and multivariate Liang–Kleeman information flow estimates diverge in the presence of conditioning variables. We therefore selected the SM/VPD→LAI/GPP couplings as the primary examples for evaluating the influence of conditioners, rather than presenting a complete bidirectional analysis of all variable pairs.

The causal graphs in Fig. 2 were obtained from the five-variable estimates including SM, VPD, T, SSR, and LAI or GPP. Accordingly, these graphs contain causal links in other directions, such as LAI→T. As explained in the final paragraph of Section 3.2, the purpose of these graphs is to illustrate the broader causal structure at selected time points characterized by large or small ΔIF . They show how the complete five-variable system behaves when the bivariate and multivariate estimates for the SM/VPD→LAI/GPP couplings strongly diverge or converge. In the selected examples, stronger divergence is associated with a richer and more interconnected causal structure, whereas stronger convergence is associated with fewer causal links. The additional arrows are therefore not intended to indicate that all causal directions are analyzed in equal detail in this manuscript.

The manuscript has been revised to clarify this point as follows:

Addition to the Methods (Section 2.1):

This directional focus reflects the roles of SM and VPD as water supply and atmospheric water demand, respectively, and the study does not aim to provide an exhaustive bidirectional analysis of all variable pairs.

Addition to the Results (Section 3.2):

The causal graphs are obtained from the corresponding five-variable multivariate systems and therefore also contain directional influences other than the SM/VPD→LAI/GPP pathways examined in detail here. These additional links are shown to illustrate the broader causal structure at the selected time points but are not analyzed individually in the present study. Further analyses will be required to systematically disentangle these additional directional connections.

Addition to the Figure 2 caption:

The causal graphs are constructed from the corresponding five-variable multivariate estimates.

4. In the discussion, Granger causality is characterized as "GC infers lagged directional dependence." Then LKIF seems like a lag-free measure. However, LKIF uses the derivative of the effect variable, computed as differenced time series. Here the monthly data are used, so one month is the effective time lag for which the causal relations are presented. Would different time lags change the results? (Larger or smaller, the data were subsampled. BTW. the way of subsampling can affect the causality results.)

Response: We thank the reviewer for this important comment. We agree that the temporal interpretation of LKIF requires clarification. In the present implementation, the derivative of the target variable is estimated through first-order Euler differencing over the monthly sampling interval ($dt=1$ month). We therefore agree that a one-month temporal step enters the LKIF estimation through the discrete approximation of the target-variable tendency. However, this is distinct from an explicit one-month lag analysis, because the source variable itself is not shifted relative to the target.

To investigate delayed relationships, LKIF can be applied explicitly at different lags by shifting the potential driving variable relative to the target and recomputing the information flow. Previous studies have shown that both lag structure and temporal resolution can influence LKIF estimates.

Docquier et al. (2023) found that introducing a one-month-lagged driver revealed information-transfer pathways that differed from the contemporaneous analysis, and that the magnitude and spatial structure of LKIF changed when moving from monthly to interannual and decadal timescales. They also showed, using synthetic experiments, that sampling frequency can affect LKIF estimation. Similarly, Docquier et al. (2024) demonstrated that explicit lagged analyses can modify the detected causal relationships.

In the present study, our objective is to investigate how conditioning variables modify LKIF estimates under a consistent monthly temporal framework rather than to identify the optimal lag structure of land–atmosphere interactions. We therefore retain the monthly temporal resolution throughout the analysis. We have now added a discussion of the potential influence of temporal averaging, sampling frequency, and explicit lag structure on the inferred information flow, and identify a systematic lag-sensitivity analysis as an important direction for future work.

The following paragraph has been added to the Discussion:

In addition to these methodological differences, the temporal resolution and lag structure used in LKIF are important for interpreting the estimated information flow. In the present study, the temporal derivative of the target variable for each information-flow direction is approximated by first-order differencing over the monthly sampling interval ($dt=1$ month). This monthly temporal step enters the LKIF estimation through the discrete approximation of the target-variable tendency but differs from an explicit lagged analysis in which the source variable is shifted relative to the target. Previous applications have shown that LKIF estimates can vary with both lag structure and temporal resolution. Introducing lagged variables can reveal causal relationships that are weak or absent in contemporaneous analyses, while temporal averaging and sampling frequency can modify the magnitude and structure of the estimated information flow (Docquier et al., 2023, 2024). Our results should therefore be interpreted within the monthly temporal framework used here. A systematic assessment of different temporal resolutions and lag structures therefore represents an important direction for future work.

Referee #5

Review of “Benchmarking Conditioners in Liang–Kleeman Information Flow: Application to Land–Atmosphere Interactions”

The revisions of the manuscript improved its clarity and limitations are now stated more explicitly. In particular, the modifications make it easier to understand the concepts used in the manuscript. However, some changes do not fully resolve the initial criticisms of the reviewers, and some new questions arise from the clarifications of the definition / interpretation of the concepts. Therefore, I recommend further revisions to the manuscript to make it suitable for publication in ESD.

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.

General comments:

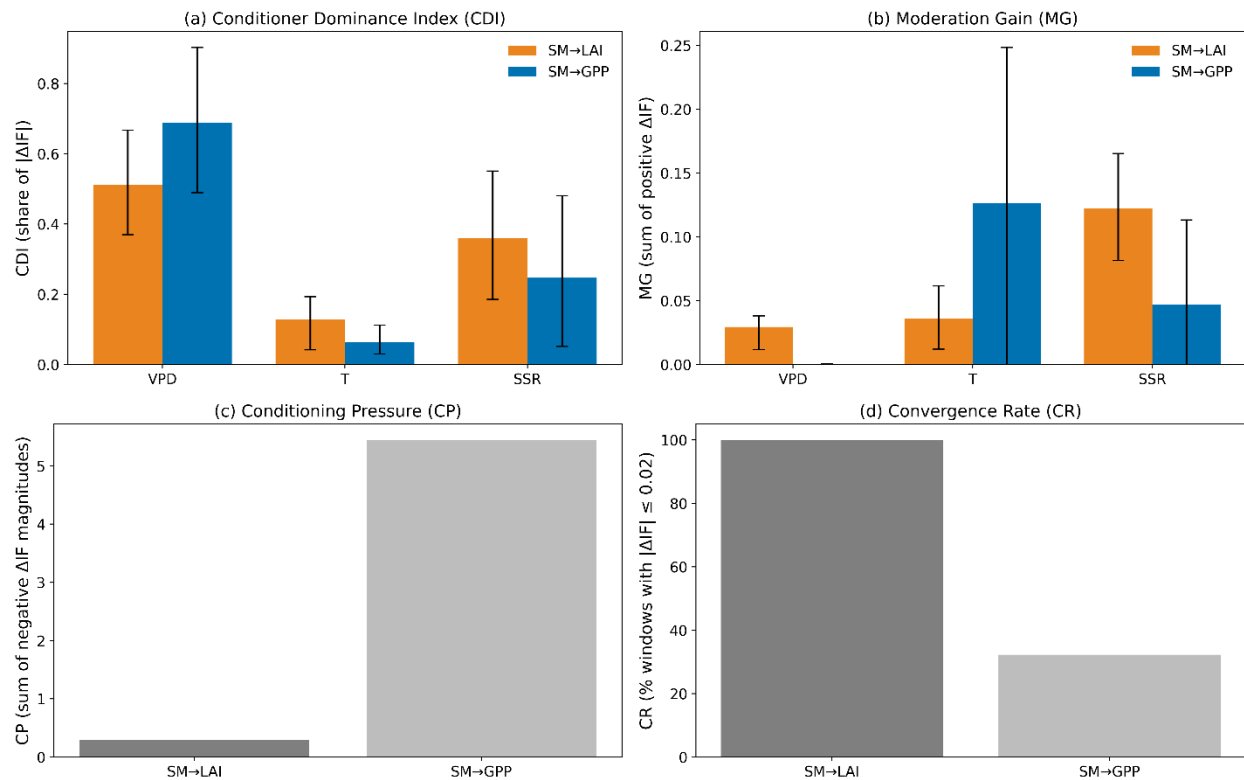
- Sect. 2.4 / Order of adding conditioners for computing the indices: It would be more convincing to report the mean and range (as uncertainty measure) over all permutations in the main manuscript instead of using one specific permutation (for example in Fig. 5).

Response: Thank you. We agree that sequential decomposition can introduce dependence on the order in which the conditioning variables are added. In the first revision, we therefore repeated the decomposition across all possible conditioner-entry permutations for each pathway, with the individual permutation results reported in Supplementary Fig. S8. To address the reviewer’s concern, we have now additionally calculated the mean and minimum–maximum range of the diagnostic indices across all permutations, and these results are presented in the new **Supplementary Fig. S9**. Both analyses are now explicitly referenced in the revised main text.

We have intentionally retained the predefined conditioner sequence used in Fig. 5 and throughout the main analysis because the primary purpose of the sequential decomposition is to track how a specific driver–response relationship changes as additional conditioners are progressively introduced under a consistently defined conditioning trajectory. Maintaining the same sequence

throughout the analysis preserves the interpretation of each successive conditioning step and allows direct comparison of the evolving relationships across space and time. An average across permutations addresses a separate question, the sensitivity of the derived indices to conditioner ordering, but it no longer represents any single sequential conditioning trajectory. Our objective is therefore pathway-specific; this is to examine how individual driver–response relationships evolve under sequential conditioning, rather than to obtain an order-independent attribution of all variables within the complete multivariate system. For this reason, we retain the predefined sequence in the main analysis and use the complete permutation analysis as a robustness assessment.

Supplementary Fig. S8 shows the results of the individual conditioner-order permutations, whereas the new Supplementary Fig. S9 summarizes their means and ranges. These analyses quantify the extent of order dependence and allow the robustness of the patterns obtained from the predefined sequence to be assessed directly. We have clarified this distinction in both the Methods and Results and now explicitly refer to Supplementary Figs. S8 and S9 when presenting the sequential decomposition in the main text.



Supplementary Fig. S9. Robustness of four diagnostic indices to conditioner ordering.

Mean values and permutation ranges of CDI, MG, CP, and CR were calculated across all possible ordering sequences of the conditioning variables (VPD, T, and SSR). Bars denote the mean across permutations and error bars indicate the corresponding range. Results are shown for both SM–LAI and SM–GPP pathways. In panel (c), CP is shown as the accumulated value across conditioners for each pathway.

- Definition and interpretation of Moderation Gain and Confounding Pressure: The authors provide some new interpretation of the indices in the revised discussion, but due to the importance of the indices, the interpretation should already be improved in the methods and results section. In particular, I find the interpretations in the methods section (and at times in the results) counterintuitive. ΔIF is positive when $T^{bi} > T^{mv}$ (Equ. 7-9). This is consistent with the figures. However, if this is the intended formulation, why do “High MG_m values identify conditioners that dampen the apparent driver influence in the bivariate signal.”? Isn’t it rather the other way around, that the bivariate signal overestimates the coupling for high MG values because it does not account for a conditioners? Similarly for the interpretation of confounding pressure: CP is positive when $T^{mv} > T^{bi}$, suggesting that the true IF (after accounting for conditioners) is higher than the estimate by T^{bi} and not that the “conditioner (m) suppresses inflated bivariate influence”? Related, I find the interpretation “Large CP_m values indicate strong confounding by the conditioner.” confusing, because (a) it is stated before that the method cannot separate between confounding and mediation, and (b) I’m more used to using confounding for cases where ignoring a third variable leads to an overestimation of the true coupling, rather than an underestimation.

Please also check the interpretations in l. 469-479 regarding the direction of the differences between bivariate and multivariate formulations. For example, Fig. 1c suggests to me that the coupling (IF) of SM $\rightarrow$ GPP is too low (dampened) when not conditioning on VPD (and T, SSR), but l. 479 states that the bivariate representation overstates the true conditioning.

Response: We thank the reviewer for identifying this inconsistency in the interpretation of the proposed indices. We agree that the previous definitions and interpretations of MG and CP were not fully consistent with the sign convention of ΔIF . We have therefore retained the mathematical formulations in Eqs. 7–12 but revised the definitions and interpretations of MG and CP throughout the Methods and Results, with corresponding revisions in the Discussion and Conclusion.

In the revised Methods, MG is defined as the positive component of ΔIF , representing decreases in IF following the inclusion of a conditioner. Accordingly, large MG_m values indicate that the IF before inclusion of conditioner m is higher than the corresponding IF after conditioning. CP represents the negative component of ΔIF , such that large CP_m values indicate increases in IF following conditioning. We have removed the previous descriptions referring to MG as dampening the apparent bivariate influence and to CP as suppressing an inflated bivariate influence. We also removed the statement that large CP necessarily indicates strong confounding, since the present framework does not distinguish between confounding and mediation.

We also revised the corresponding interpretation in the Results. Importantly, Eqs. 7–9 describe a sequential conditioning procedure. Thus, the conditioner-specific MG and CP components represent changes in IF relative to the preceding conditioning step rather than, for every conditioner, a direct comparison between the original bivariate and fully conditioned estimates. In Fig. 5c, CP is shown as the accumulated value across conditioners for each pathway. This distinction is now made clear in the interpretation of Fig. 5. For example, the large MG associated with temperature for $SM \rightarrow GPP$ is now interpreted as the largest decrease in IF when temperature is included as a conditioner, relative to the preceding conditioning step.

We further corrected the directional interpretation of CP in the Results. For $SM \rightarrow GPP$, the large accumulated CP is now interpreted as reflecting increases in IF across the conditioning sequence. Consistent with the reviewer's observation of Fig. 1c, we now state explicitly that the bivariate $SM \rightarrow GPP$ estimate is generally lower than the corresponding fully conditioned estimate. The previous statement that the bivariate representation “overstates” the conditioned coupling has therefore been removed and replaced with the correct directional interpretation.

Finally, we revised the corresponding wording in the Conclusion so that it is consistent with the corrected MG interpretation. These revisions ensure that the definitions, interpretation of the Results, figures, and Conclusion are consistent with the sign convention used in Eqs. 7–12.

The corresponding text in the Methods (Sections 2.4.2 and 2.4.3) has been revised as follows:

2.4.2 Moderation Gain (MG)

Moderation Gain (MG) shows how strongly the IF decreases after including the conditioner (m). The gain is defined as (Eq. 11);

$$MG_m = \sum_t \max(0, \Delta IF_m(t)), \quad (11)$$

High MG_m values indicate that the IF before inclusion of the conditioner is higher than the corresponding IF after conditioning.

2.4.3 Conditioning Pressure (CP)

Conditioning Pressure (CP) measures how strongly the IF increases after including conditioner (m) and is given in Eq. 12;

$$CP_m = \sum_t \max(0, -\Delta IF_m(t)) \quad (12)$$

Large CP_m values indicate that conditioning increases the inferred IF.

The corresponding text in the Results (Section 3.5) has been revised as follows:

MG (Fig. 5b) captures the sum of positive ΔIF values, representing decreases in IF across the conditioning sequence. For SM→LAI, MG is small for all conditioners, increasing slightly from VPD to T and reaching its largest value for SSR. This pattern indicates that T and SSR contribute larger decreases in IF than VPD, although these changes remain small for this pathway. In contrast, SM→GPP exhibits much larger MG dominated by temperature, with T producing a considerably larger moderation gain than VPD or SSR. This indicates that T contributes the largest decrease in IF relative to the preceding conditioning step, while SSR contributes a smaller decrease and VPD contributes little to the positive ΔIF component.

CP (Fig. 5c) measures the accumulated negative ΔIF and identifies cases in which conditioning increases the inferred IF. SM→LAI displays very low CP, indicating little increase in IF after conditioning. SM→GPP, however, shows much larger accumulated CP, indicating that conditioning produces a larger increase in the inferred IF. This is consistent with Fig. 1c, where the bivariate SM→GPP estimate is generally lower than the corresponding conditioned estimates. This confirms that the bivariate representation of SM→GPP can markedly understate the inferred IF relative to the conditioned coupling.

• Sect. 3.4 / Interpretation of the ANOVA: Strictly speaking, the ANOVA doesn't identify specific values of drivers/conditioners that lead to large $|\Delta IF|$ values. It only identifies which (combination of) variable(s) can explain the variability of $|\Delta IF|$. While the interpretations in the section are ecologically plausible, a different model is required to quantify the direction and effect sizes of these influences on $|\Delta IF|$, for example a multiple linear regression model. Please also comment on the very strong influence of the three-way interactions on the explained variability, as according to Fig. S4 the direct influences and two-way interactions seem to explain essentially no $|\Delta IF|$ variability.

Response: We thank the reviewer for this helpful and important clarification. We agree that the ANOVA variance decomposition does not, by itself, identify the particular quartile levels that produce large $|\Delta IF|$, nor does it estimate the direction of the association between each environmental variable and $|\Delta IF|$. Its purpose in our analysis was narrower: to determine whether variability in $|\Delta IF|$ is primarily associated with individual hydroclimatic factors or with their interactions. (Figure S4 is now S5)

We selected a factorial ANOVA because our primary interest was in regime dependence and non-additive interactions among moisture supply, atmospheric demand, and energy availability. Categorizing the variables into quartiles allowed us to compare physically interpretable hydroclimatic regimes without imposing a monotonic linear relationship between the continuous drivers and $|\Delta IF|$. The resulting ANOVA quantities, including the variance-based effect size ω^2 and explained SD, should not be interpreted as regression slopes or as evidence that a particular factor level increases or decreases $|\Delta IF|$.

We have revised Sect. 3.4 accordingly. Statements identifying specific dry, hot, or high-radiation conditions as producing the largest departures have been removed, together with ecological interpretations that were not directly supported by the variance partitioning. Directional or level-specific effects would require additional analysis, such as appropriate contrasts of the factorial model or a complementary regression framework. Because our objective here is variance attribution and interaction structure rather than estimation of directional effects, we did not introduce an additional regression model.

We have also revised the discussion of the three-way interactions. In the three-factor analyses, the $SM \times VPD \times T$ and $SM \times VPD \times SSR$ interactions are the dominant components of the variance partitioning, whereas the main effects and two-way interactions contribute comparatively little. We now interpret this pattern as a strongly non-additive dependence of $|\Delta IF|$ on the joint hydroclimatic regimes, rather than as evidence that any individual environmental factor consistently increases or decreases $|\Delta IF|$. We have accordingly removed ecological interpretations that were not directly supported by the ANOVA variance partitioning.

The text in the Results (Section 3.4) has been revised as follows:

3.4 ANOVA of Regime-dependent Changes in Information Flow

This ANOVA framework (Appendix B) provides a variance-partitioning approach to quantify how variability in $|\Delta IF|$ is associated with individual hydroclimatic factors and their interactions.

a. Two-way ANOVA: moisture–demand control of $|\Delta IF|$

Across all four pathways ($SM \rightarrow LAI$, $VPD \rightarrow LAI$, $SM \rightarrow GPP$, $VPD \rightarrow GPP$), the two-factor ANOVA partitions variability in $|\Delta IF|$ among the main effects of SM and VPD and their interaction (Fig. 4a). For the SM-driven pathways, the SM main effect dominates, whereas the VPD main effect contributes very little to $|\Delta IF|$ for these pathways.

For VPD-driven pathways, the $SM \times VPD$ interaction is the only dominant component, showing that the magnitude of the difference between the bivariate and multivariate IF estimates is more strongly associated with the combined moisture–demand regime rather than with either factor independently. The explained SD values are generally smaller for the LAI pathways and substantially larger for the GPP pathways.

b. Thermal modulation: $SM \times VPD \times T$ structure

When temperature is included as a third factor, the $SM \times VPD \times T$ interaction becomes the dominant component of the variance partitioning across all four pathways (Fig. 4b), whereas the main effects and two-way interactions contribute comparatively little. This dominance indicates a strongly non-

additive relationship, in which variability in $|\Delta IF|$ is associated with the joint SM–VPD–T regimes rather than with consistent marginal associations of SM, VPD, or T individually.

c. Radiative modulation: SM×VPD×SSR structure

When SSR is included as a third factor, a similar pattern emerges, with the SM×VPD×SSR interaction becoming the dominant component of the variance partitioning across all four pathways (Fig. 4c), while the main effects and two-way interactions contribute comparatively little. This indicates that variability in $|\Delta IF|$ is primarily associated with the joint SM–VPD–SSR regimes rather than with consistent marginal associations of SM, VPD, or SSR individually. The corresponding variance fractions (ω^2) for each pathway are shown in Supplementary Fig. S5.

Altogether, the ANOVA results show that variability in $|\Delta IF|$ is strongly associated with interactions among the hydroclimatic factors, with the three-way interactions dominating when T or SSR is included. Although the explained SD values are substantially larger for the GPP pathways than for LAI (Fig. 4), the relative variance fractions (ω^2) show broadly similar patterns across pathways (Supplementary Fig. S5). The larger explained SD for GPP therefore largely reflects its larger overall $|\Delta IF|$ variability and does not imply a correspondingly greater relative contribution of the individual factors or interactions. Overall, these results indicate that variation in $|\Delta IF|$ depends strongly on the joint hydroclimatic regime, rather than on consistent effects of individual factors alone. Sensitivity analyses based on tercile- and quintile-based partitions show that the dominant higher-order interaction structure remains consistent (Supplementary Fig. S5). Additional details on the ANOVA formulation and robustness considerations are provided in Appendix B.

- Mediation/confounding/conditioning: conditioning is still not used consistently across the manuscript instead of either confounding or mediation, since according to the authors the LKIF cannot separate distinguish between the two (this includes the index name “confounding pressure”)

Response: We appreciate this feedback. To ensure consistency, we have standardized our terminology to conditioning throughout the manuscript. Additionally, we have renamed the index “confounding pressure” as “Conditioning Pressure” and the “Mediator Dominance Index” as the “Conditioner Dominance Index” throughout the revised manuscript. These terminology changes

have also been applied consistently in all figures and tables in both the main manuscript and the Supplementary Material.

Specific comments

- 1. 63-65: This sentence is still unnecessary vague and difficult to understand. It would be more informative to describe the main ideas / axioms behind the formalism. I also don't understand what the "contrast" is between the first and the second part of the sentence. While some other methods are mentioned in the revised manuscript, it would be helpful for readers to state more explicitly what makes the LKIF approach different from other causal inference methods and what limitations it overcomes.

Response: We thank the reviewer for this helpful comment. We agree that our previous sentence did not explain sufficiently what is meant by "derived from first principles," and that the intended contrast was unclear. We have therefore expanded this passage. We respectfully retain the first-principles distinction, however, because it is a central theoretical feature of the LKIF framework rather than a general statement that all causal methods are equivalent in their foundations.

In LKIF, the variables are represented as components of an underlying deterministic or stochastic dynamical system. The directed information flow ($T_{X \rightarrow Y}$) is obtained by decomposing the rate of change of the marginal Shannon entropy of (Y) into the contribution produced by its evolution when the influence of (X) is instantaneously excluded and the contribution transferred from (X). This decomposition is derived from the evolution of the system's probability density, using the Frobenius–Perron or Liouville/Fokker–Planck framework as appropriate, rather than postulating an information-transfer statistic empirically.

An important consequence is the principle of nil causality: if the evolution law of (Y) is independent of (X), then the information flow ($T_{X \rightarrow Y}$) is identically zero. In the LKIF framework, this property is established as a theorem rather than introduced as an axiom or assessed only through numerical validation.

We have also made the comparison with other methods more explicit. Correlation is symmetric and therefore cannot establish directionality. Granger causality and transfer entropy define causal

influence through predictive improvement or conditional statistical dependence. PCMCI reconstructs a causal graph by iteratively testing conditional independence among lagged variables. LKIF instead quantifies the signed rate at which one component of a dynamical system contributes to the entropy tendency of another. Docquier et al. (2024; doi: 10.5194/npg-31-115-2024) discuss this distinction directly in their comparison of LKIF and PCMCI.

We have nevertheless avoided presenting LKIF as universally superior or assumption-free. Although the underlying information-flow theory is derived for general deterministic and stochastic dynamical systems, the estimator employed in the present study uses a specified dynamical representation and remains sensitive to the adequacy of the observed variable set, sampling interval, data length, and estimation assumptions. We have added these qualifications to the revised manuscript.

We have revised the text as follows:

The LKIF framework differs from causal measures introduced primarily through statistical association, prediction, or conditional-independence criteria because its information-flow rate is derived from the entropy evolution of an underlying dynamical system (Liang, 2016). Specifically, the information flow ($T_{X \rightarrow Y}$) is obtained by decomposing the rate of change of the marginal entropy of Y into the entropy tendency arising when the influence of X is instantaneously excluded and the contribution transferred from X . This derivation yields the principle of nil causality as a theorem: if the evolution law of Y is independent of X , then $T_{X \rightarrow Y} = 0$.

In contrast, Granger causality and transfer entropy define influence through improved prediction, while PCMCI identifies causal links through iterative conditional-independence testing among lagged variables. LKIF provides a directional, signed rate of entropy transfer grounded in dynamical systems rather than only a measure of statistical dependence (Liang, 2016; Docquier et al., 2024). Its time-varying and multivariate extensions allow this rate to evolve through time and to be estimated while accounting for other observed variables. Nevertheless, the empirical estimator remains dependent on the assumed dynamical representation, sampling characteristics, and adequacy of the variables included in the analysis.

• 1. 124-135: Is this the same specification of the Kalman Filter as in Z24? What is the rationale for the window lengths (24 months and 90 months)? Since the optimal specification is likely depending on the length of the timestep, it would be helpful here to rather state the lengths in months and not in timesteps. In this context I also wonder how IF is estimated for the years 2004 - 2010 if date is used from 2004-2018 but the Kalman filter uses a 90 months look-back window. Are shorter windows used for the first years? If yes, could that explain why IF estimates are mostly near 0 at the start of the analysis period?

Response: Thank you for the comment. We clarify that the Kalman filter configuration used in this study is not identical to that used in Z24. Rather, our approach builds upon the framework introduced by Zhou et al. (2024) and further extends it to explore the role of conditioning variables in reconstructing causal relationships. Z24 did not consider some of the conditioning variables used in the present study, as their objective was not to quantify the individual influence of different conditioners. In contrast, our study specifically investigates how the inclusion of different conditioning variables modifies the inferred causal structure. Our results demonstrate that, even for the same variables and datasets, conditioners can reconstruct causal relationships differently across different time periods and spatial locations. Therefore, to better capture these evolving impacts of conditioning variables, the Kalman filter parameters were retuned for this specific application. Different window lengths have been adopted in previous time-varying LKIF studies depending on the research objectives and data characteristics (e.g., Hagan et al., (2019: doi:10.1175/JCLI-D-18-0881.1); Zhang et al., (2026: doi:10.21203/rs.3.rs-9450690/v1)). During the optimization process, we tested different window lengths and selected the one that provided the most scientifically interpretable results for evaluating the impact of conditioning variables. Following the reviewer's suggestion, we have revised the manuscript to report the window lengths in months rather than timesteps.

We also want to clarify that the IF estimation was not performed using only the 2004–2018 period. The vegetation datasets used in this study extend from 1981 to 2018 for LAI and from 1982 to 2018 for GPP. Therefore, the analysis was performed using the full available time series (1981/82-2018). However, considering uncertainties associated with the early satellite products and the selected window length for the Kalman filter, we restricted the interpretation and presentation of

the results to the period 2004–2018. We have revised the data section to clarify this distinction between the full analysis period and the period used for interpretation.

The Data section now reads as follows:

2.5 Data

The Gross Primary Productivity (GPP) data, reflecting photosynthetic activity of plants, and Leaf Area Index (LAI), representing vegetation greenness, utilized in this study, are obtained from the Global Land Surface Satellite (GLASS) dataset (Liang et al., 2021). Both the LAI and GPP datasets cover the period 1981–2018 and have a spatial resolution of 0.05° and a temporal resolution of 8 days. For consistency with other variables in the analysis, the datasets are resampled to a spatial resolution of $0.25^\circ \times 0.25^\circ$ (longitude $\times$ latitude) using bilinear interpolation and then aggregated to monthly timescales.

The monthly data on Volumetric Soil Moisture (SM), 2-m air temperature, and Surface Solar Radiation (SSR) used in this study are obtained from the European Centre for Medium-Range Weather Forecasts' (ECMWF) datasets, with SM taken from ERA5-Land and temperature and SSR from ERA5 global atmospheric reanalysis products. The weighted average of SM, computed between 0 and 289 cm (all four layers of SM), is used here, with the weights corresponding to the thickness of each layer (0–7, 7–28, 28–100, and 100–289 cm). The original resolution of the SM data ($0.1^\circ \times 0.1^\circ$) is subsequently re-gridded to a horizontal resolution of $0.25^\circ \times 0.25^\circ$. Vapor pressure deficit (VPD) is estimated using ERA5's 2-m minimum and maximum temperature and 2-m dewpoint temperature (Upreti & Ojha, 2017). The monthly anomalies for all variables are computed over the corresponding analysis period by subtracting the respective climatological means from the original data. To ensure temporal consistency with the vegetation records, these variables are used over the period from 1981 to 2018.

LKIF estimation is performed from 1981 to 2018 for both the LAI- and GPP-based analyses. However, given the uncertainties associated with the earlier vegetation products and to ensure adequate preceding observations for the Kalman filter estimation, only the results from 2004 to 2018 are presented and interpreted in this study.

- 1. 145-152: This is another example where more work to understand the interpretation of the methods would be very helpful because the description of the meaning of the IF stays very vague

and hard to grasp. In particular “the likelihood of control” is rarely mentioned afterwards, but the IF is rather interpreted as a strength of control and not a likelihood. It would also be helpful to state more explicitly what is meant by increased vulnerability.

Response: We thank the reviewer for this helpful comment. We agree that our previous use of the term “likelihood of control” was inappropriate and potentially misleading. Although the parameters entering the LKIF estimator are obtained through maximum-likelihood estimation, the resulting information flow (IF) is not itself a likelihood or probability. We have therefore removed this terminology and revised the manuscript to describe IF as a signed rate of directed information transfer from a source variable X to a target variable Y . Within the LKIF framework, a statistically significant nonzero $IF_{X \rightarrow Y}$ indicates a directed influence of X on Y , while the magnitude $|IF_{X \rightarrow Y}|$ characterizes the strength of the inferred directed influence. This interpretation is consistent with the nil-causality property of LKIF, whereby the information flow vanishes when the evolution of the target is independent of the source.

We have also clarified the interpretation of the sign of IF. Importantly, the sign does not indicate whether the target variable itself increases or decreases. Rather, positive IF indicates that the source exerts a variability-enhancing influence on the target, whereas negative IF indicates a variability-reducing influence. This interpretation follows the information-theoretic meaning of LKIF and is consistent with Docquier et al. (2024). In particular, Docquier et al. (2024) explicitly distinguish the magnitude of LKIF, which reflects the strength of the causal influence, from its sign, for which positive and negative values indicate increasing and decreasing target variability, respectively.

We agree with the reviewer that our previous use of “vulnerability” was insufficiently defined and too directly associated with the sign of IF. Our original motivation came from the broader ecological literature linking information-theoretic concepts with ecosystem stability, including Rutledge et al. (1976, doi: 10.1016/0022-5193(76)90007-2). However, this literature does not establish a formal equivalence between positive/negative LKIF and ecological vulnerability/resilience. We have therefore removed the general interpretation of +IF as vulnerability and -IF as resilience from the methodological definition.

In the revised manuscript, the terms vulnerability and resilience are used only when an ecological interpretation is independently supported by the physical context. In such cases, “increased vulnerability” refers to a greater susceptibility of vegetation structure or function to adverse

hydroclimatic perturbations, rather than to a mathematical property of positive IF itself. Otherwise, we restrict the interpretation to the methodologically supported terms variability-enhancing and variability-reducing influence.

The interpretation of Information Flow in Section 2.1 has been revised as follows:

We interpret LKIF as a signed rate of directed information transfer from a source variable to a target variable. A statistically significant nonzero LKIF indicates a directed causal influence of the source on the target within the LKIF framework, while the absolute magnitude of LKIF characterizes the strength of the directed influence. When LKIF is positive (+IF), this suggests that variability in the source increases the variability of the target. In contrast, a negative LKIF (−IF) value indicates that variability in the source decreases the variability of the target (Liang, 2014; Docquier et al., 2024).

• 1. 178-180: Does that preserve the spatial (and/or temporal) autocorrelation structure of the data? This would be necessary for the bootstrap to provide meaningfully superior results to the t-test which assumes independence of the samples.

Response: Thank you. We agree that the spatial dependence among neighboring grid cells needs to be considered when using the spatial ensemble for resampling. The bootstrap used here is implemented as a paired moving spatial block bootstrap at each monthly time step. Contiguous 3×3 grid-cell blocks are resampled with replacement while preserving the pairing between the bivariate and multivariate IF estimates. At the 0.25° spatial resolution used in this study, this corresponds to an approximately 0.75°×0.75° spatial neighborhood. This procedure retains local spatial dependence among neighboring grid cells rather than treating individual cells as fully independent resampling units.

The 95% confidence intervals obtained using the 3×3 spatial block bootstrap identify the same significant periods as the paired t-test shown in Fig. 2. We emphasize that the bootstrap is used here as a complementary sensitivity assessment rather than as a replacement for the paired t-test.

We also examined the sensitivity of the bootstrap confidence intervals to larger spatial block sizes. As expected, increasing the block size produces wider confidence intervals and can lead to a loss of significance for some months close to the zero threshold, because larger blocks reduce the

effective number of spatial resampling units. This indicates that the exact pointwise significance classification can depend to some extent on the assumed spatial dependence scale. We therefore interpret the bootstrap confidence intervals primarily as regional uncertainty bounds rather than exact pointwise significance tests.

Regarding temporal autocorrelation, the bootstrap is performed separately at each monthly time step, with the spatial ensemble of grid cells providing the resampling dimension. Temporally successive observations are therefore not treated as independent replicates within each month-specific bootstrap. Nevertheless, successive IF estimates may retain temporal dependence because they are derived from temporally filtered series, and we have clarified this qualification in the revised manuscript.

We have made this clear in the revised manuscript as follows:

To evaluate the sensitivity of this significance assessment to the choice of testing framework, we additionally performed a paired spatial block bootstrap at each monthly time step. The bivariate–multivariate IF differences were resampled using contiguous 3×3 grid-cell blocks, while preserving the pairing between the two estimates. This block resampling retains local spatial dependence among neighboring grid cells rather than treating individual cells as fully independent resampling units. The bootstrap-based 95% confidence intervals identified the same significant periods as the paired t-test, and the bootstrap is therefore used here as a complementary sensitivity assessment of the t-test-based significance pattern.

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 .

The bootstrap is performed separately at each monthly time step, with the spatial ensemble of grid cells providing the resampling dimension; temporally successive observations are therefore not treated as independent bootstrap replicates. Nevertheless, successive IF estimates may retain temporal dependence because they are derived from temporally filtered series. In addition, spatial

block resampling retains local dependence among neighboring grid cells but may not fully capture longer-range spatial dependence. And the resulting confidence intervals may show some sensitivity to the assumed block size. These confidence intervals should therefore be interpreted as approximate regional uncertainty bounds rather than exact pointwise significance tests.

• l. 249: How is $\varepsilon = 0.02$ selected? What percentile across all couplings is that (since this is the proposed threshold selection criterion in the brackets above)?

Response: Thank you. The value $\varepsilon = 0.02$ corresponds to the 10th percentile across all couplings. This is already mentioned in the manuscript in Section 2.4.4 (lines 247-248; now 263-264 in the revised version).

• l. 291: I find the reformulation of this sentence still unnecessarily difficult to understand. Doesn't the loss of significance simply indicate an overestimation of the directional coupling when not accounting for conditioners.

Response: We agree with the reviewer. We have revised the sentence, and the new sentence reads as "The loss of significance indicates an overestimation of the directional coupling in the bivariate form".

• l. 332-333: Does that indicate that conditioning improves the IF estimation in this situation and identifies a bias of the bivariate framework? It would be nice to state this more explicitly if this is meant.

Response: Thank you. That is indeed what we want to convey. We have revised the sentence to state as, "This sensitivity explains why conditioning on these variables changes the apparent strength of VPD's impact, thereby correcting bivariate bias and providing a more refined IF estimate".

• l. 350: What is meant by "feedback" here and is there a difference between coupling and feedback?

Response: We thank the reviewer for this question. To clarify, the phrase "couplings/feedback" appears just before line 350 (lines 345–347; now 373-374 in the revised version) in Section 3.2.

For the distinction between coupling and feedback, we refer to Appendix A of Seneviratne et al. (2010, doi:10.1016/j.earscirev.2010.02.004).

Following this framework;

1. Coupling denotes a one-way directional control from a driver to a target variable (e.g., $SM \rightarrow GPP$).
2. Feedback refers to a two-way, bidirectional coupling where variables mutually influence each other (e.g., $SM \leftrightarrow GPP$).

In our study, feedback refers to the two-way coupling between climate variables and vegetation (e.g., $SM \leftrightarrow VPD$ and $VPD \leftrightarrow T$).

• 1. 448-451: This difference seems to be only because of the larger overall variability of $|\Delta IF|$ for GPP than LAI. The relative explained variances of the factors are very similar (see Fig. S4).

Response: Thank you, we agree with the reviewer. We have revised Sect. 3.4 to clarify that the larger explained SD values for the GPP pathways mainly reflect the larger overall variability in $|\Delta IF|$, while the relative variance fractions (ω^2) show broadly similar patterns across pathways (now Supplementary Fig. S5). We have accordingly removed the previous interpretation of the larger GPP explained SD as evidence of stronger regime dependence than for LAI. Further details on the revised interpretation of the results are provided in our response to [General Comment 03](#).

• 1. 469: I don't think "amplify" is a good wording here, because MG and CP are measuring opposing directions by which bivariate and multivariate estimates differ, but not opposing amplitudes. Please choose a more careful wording. If you indeed want to measure amplification / dampening, I would rather define MG and CP as multiplicative measures and not additive measures.

Response: We thank the reviewer for this helpful distinction. We agree that "amplify" was not appropriate in the context of MG and CP. As presently defined, these indices separate the positive and negative components of the additive difference between the bivariate and conditioned IF estimates. They therefore quantify opposing directions of conditioning-induced adjustment, rather than opposing amplitudes or multiplicative amplification and dampening.

This interpretation is distinct from the meaning of the sign of IF itself. Within the LKIF framework, positive IF indicates that the source contributes to increasing the marginal entropy of the target

and therefore represents a variability-enhancing contribution. Negative IF indicates that the source contributes to reducing the target's marginal entropy and therefore represents a variability-reducing contribution. We retain this information-theoretic interpretation when discussing the sign of IF, but no longer use “amplify” or “dampen” to characterize MG and CP.

We have retained the additive definitions because our purpose is to summarize the cumulative direction and magnitude of the adjustment introduced by conditioning, not to estimate an amplification factor. A multiplicative formulation would answer a different question and would become unstable or undefined when IF is close to zero or changes sign. We have therefore revised the manuscript to describe MG and CP as cumulative positive and negative conditioning adjustments between successive IF estimates, without interpreting them as measures of amplification or attenuation.

The revised text is below (Section 3.5):

MG (Fig. 5b) captures the sum of positive ΔIF values, representing decreases in IF across the conditioning sequence. For SM→LAI, MG is small for all conditioners, increasing slightly from VPD to T and reaching its largest value for SSR. This pattern indicates that T and SSR contribute larger decreases in IF than VPD, although these changes remain small for this pathway. In contrast, SM→GPP exhibits much larger MG dominated by temperature, with T producing a considerably larger moderation gain than VPD or SSR. This indicates that T contributes the largest decrease in IF relative to the preceding conditioning step, while SSR contributes a smaller decrease and VPD contributes little to the positive ΔIF component.

CP (Fig. 5c) measures the accumulated negative ΔIF and identifies cases in which conditioning increases the inferred IF. SM→LAI displays very low CP, indicating little increase in IF after conditioning. SM→GPP, however, shows much larger accumulated CP, indicating that conditioning produces a larger increase in the inferred IF. This is consistent with Fig. 1c, where the bivariate SM→GPP estimate is generally lower than the corresponding conditioned estimates. This confirms that the bivariate representation of SM→GPP can markedly understate the inferred IF relative to the conditioned coupling.

• Fig. 5c: CP is defined in the methods for individual conditioners, but no individual conditioner is mentioned here. Why?

Response: We thank the reviewer for pointing this out. In Fig. 5c, CP represents the accumulated conditioning pressure across all conditioners rather than the contribution of an individual one. The purpose of this panel is to show the overall effect of conditioning on the bivariate estimate. This was already described in the text around line 475 (499 in revised version), where CP is presented as the accumulated negative ΔIF contribution. To make this clearer, we have added the following statement to the figure caption: “In panel (c), CP is shown as the accumulated value across conditioners for each pathway.”

- 1. 595: Please rephrase to “Synthetic data experiments” instead of “Theoretical experiments” because this is just one example with synthetic data and not a theoretical proof.

Response: Thank you. We have revised the phrasing to “Synthetic data experiments”.