Responses to Referee #2 comments:

Manuscript number: **egusphere-2025-4213**

Title: **Is earlier always better? A comparative assessment of an early-stage response window for rainfall timing in multiyear drought alleviation in China**

We sincerely thank you for the valuable comments and suggestions, which have greatly improved our manuscript. We have addressed all the comments and made the revisions. Our responses to each comment are in blue text, while the referees' comments are in black. For clarity and ease of cross-reference, all figure and table numbers cited in this response correspond directly to the numbering used in the revised manuscript and Supplementary Material.

General Comment

I would like to thank the authors for revisiting their manuscript according to the reviewers' suggestions, and to provide a much-needed clarification on the motivation behind the study. Unfortunately, beside this clarification, I do not see any major improvement of the manuscript, as the results remain mostly unchanged and the major critical points untouched.In the current status, the manuscript does not go beyond a simple numerical exercise without a proper connection to the real world. In addition to this limitation, the obtained results are strongly constrained by the simplified hypotheses adopted for its implementation, providing little generality. It follows that the scientific question "is early always better?" remains unanswered, as the differences observed in the reported results remain well-within the broad range of uncertainty.

Response: We sincerely thank you for this critical and constructive comment. In the revised manuscript, these concerns have been carefully considered and addressed through a series of targeted revisions, including additional robustness evaluations, statistical analyses, and expanded process-based interpretations. The corresponding details are further discussed in the following responses.

Regarding the connection to real-world processes, the analysis is based on long-term observational records from meteorological stations and employs the widely used Palmer Drought Severity Index (PDSI) to characterize drought conditions. This ensures that the numerical experiment is grounded in observed data and reflects realistic drought dynamics. In addition, multiyear drought (MYD) events are specifically targeted because their cumulative water deficits over extended periods make them particularly suitable for examining precipitation effects on drought severity.

Concerning the use of simplified assumptions, the numerical experiment was designed to provide a controlled framework for examining the influence of a single factor, namely precipitation. Within this framework, the sensitivity of MYD events to hypothetical precipitation inputs applied at different timings is quantitatively assessed using the Drought Attenuation Quantitative Model (DAQM). To further evaluate the robustness of the DAQM-based results, additional empirical estimation methods were introduced to identify the optimal timing for MYD events alleviation ($t_{optimal}$) **(Pages 14 to 15, Lines 322 to 362 and Pages 23 to 24, Lines 512 to 535)**. During the numerical experiments, it was found that the identified $t_{optimal}$ for a given MYD event was not always identical across different hypothetical precipitation perturbation scenarios, indicating that precipitation perturbation magnitude itself can influence the identified $t_{optimal}$. To reduce the influence of perturbation magnitude on $t_{optimal}$

identification, 21 hypothetical precipitation perturbation scenarios (pr_{rate}=0, 0.05, 0.10, ..., 1.00) were separately evaluated for each MYD event, thereby integrating both small and relatively large perturbations into a comprehensive evaluation method. For each scenario, the corresponding optimal timing was first determined according to the maximum DS_{rate} among different $t_{forcing}$. The mode of the resulting 20 optimal timings was then used to represent $t_{optimal}$ across different precipitation perturbation magnitudes (within the mode-based identification process, scenario without additional precipitation input, i.e., pr_{rate}=0, was not included because identical DS_{rate} across all $t_{forcing}$ prevented a unique optimal timing from being identified), hereafter referred to as the mode-based approach. Because $t_{optimal}$ is expressed as a discrete positive integer, the mode was adopted instead of the mean or median to ensure that the identified $t_{optimal}$ remained physically interpretable as an integer month. Notably, the DAQM-based estimates are also derived from all hypothetical precipitation perturbation scenarios through least-squares fitting, making the mode-based and DAQM-derived results directly comparable **(Page 14, Lines 322 to 342)**. Generally, small perturbation experiments (pr_{rate} = 0.05) represent more frequently occurring weak perturbation conditions in real-world situations. Based on this consideration, a small-perturbation experiment (pr_{rate}= 0.05) was conducted as an independent robustness check to evaluate whether the identified $t_{optimal}$ remains stable under small precipitation perturbations **(Pages 14 to 15, Lines 343 to 359)**. The results show that this framework captures MYD behavior well at the grid-level event scale, with relatively high agreement among the DAQM-derived results, the mode-based results, and the small perturbation-based results **(Pages 14 to 15, Lines 322 to 362 and Pages 23 to 24, Lines 512 to 535)**. In this context, the results provide theoretical support for understanding the role of precipitation timing in drought alleviation.

To address the concern regarding uncertainty and the robustness of identified timing differences, the revised manuscript now incorporates additional statistical significance tests **(Page 12, Lines 289 to 301)**, Cohen's *d* analyses **(Page 12, Lines 289 to 301, and Page 13, lines 311 to 316)**, comparison among DAQM-derived, mode-based, and small precipitation perturbation-based $t_{optimal}$ **(Pages 14 to 15, Lines 322 to 362 and Pages 23 to 24, Lines 512 to 535)**. These analyses consistently show that, although the differences among t_1, t_2, and t_3 are relatively small when all grid-level MYD events across the study area are considered collectively, they remain distinguishable at the river basin scale **(Fig. S4)** and spatially robust. Additional details regarding model accuracy, forcing uncertainty, spatial aggregation, and the interpretation of differences among t_1~t_3 are provided in the following responses.

The revised manuscript also further clarifies the interpretation of the central question "is earlier always better?". The results do not suggest that later precipitation input is generally more effective than earlier precipitation input. Instead, the revised analyses reveal the existence of an early-stage response window during MYD evolution, corresponding primarily to t_1~t_3. Within this window, attenuation efficiency is jointly influenced by the cumulative temporal advantage associated with earlier precipitation input and the higher numerical sensitivity of standardized drought indices under more severe drought conditions **(Page 22, Lines 455 to 484)**. Consequently, although t_1 remains the dominant optimal timing overall, precipitation input during t_2 or t_3 may still exhibit attenuation efficiencies comparable to, or occasionally exceeding, that of t_1 under certain conditions. Importantly, even within this early-stage response window, a non-negligible proportion of cumulative deficit has already developed **(Fig. 4b)**. This indicates that the first three months collectively represent a critical stage for MYD alleviation. From a practical perspective, these results suggest that missing t_1 does not necessarily imply that the

opportunity for relatively high attenuation efficiency has already been lost **(Pages 22 to 23, Lines 462 to 511)**.

These aspects have been clarified and incorporated throughout the revised manuscript.

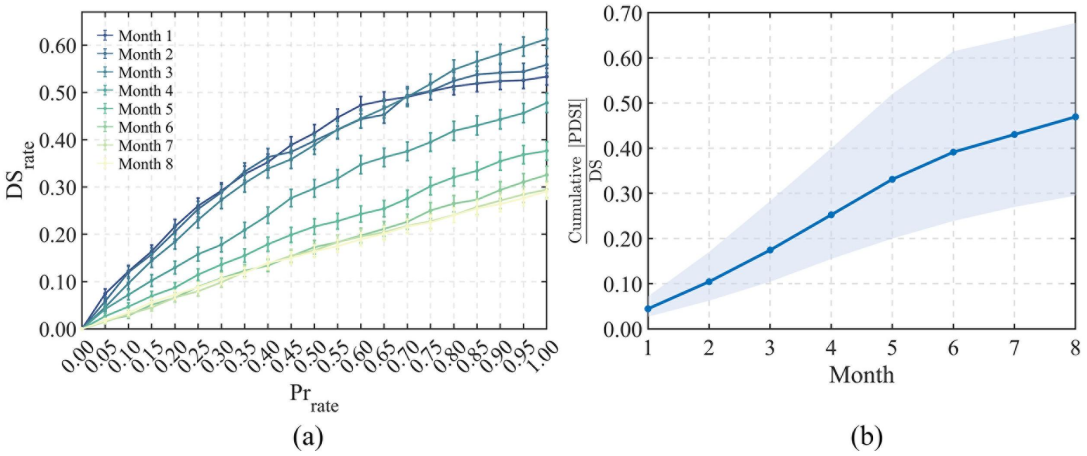


Figure 4: DS responses to hypothetical precipitation perturbations and cumulative absolute PDSI from event onset to a given month within a single event relative to the DS of that event during DDP in China. (a) Mean responses under different hypothetical precipitation perturbations for t_1~t_8 across all MYD events, with error bars representing standard error. (b) Temporal evolution of $\frac{cumulative\ |PDSI|}{DS}$ during DDP across all MYD events.

In particular, the authors adopt an exponential relationship to approximate the attenuation effect and select the "optimal" timing based on the maximum value of the "k" factor. They show that over about 33% of the domain, t2 and t3 are selected over t1. However, the obtained results are not contextualized in contrast to: model accuracy, significance of the differences, uncertainty in the adopted forcings, spatial aggregation, etc.
Just as an example, there is a row of grid cells in the north-west where the t2 is the optimal choice (Fig. 6b). In this areas, relative errors are in the order of 5-6% (Fig. 6a), and differences t1-t2 of about 0.06 (Fig. 7b). Are the results between t1 and t2 different enough to justify the statement that "t2 is a better optimal than t1"? How does this compare with uncertainty in precipitation? Can a small change in the forcing causing a similar difference? Differences in Fig. 7 show that when t2/t3 are larger than t1, this is often with small differences compared to when the opposite happens (colour scale goes from -0.24 to 0.08, 3 times larger).

Response: We sincerely thank you for highlighting these important aspects regarding model accuracy, significance of the differences, uncertainty in the adopted forcings, and the spatial aggregation. These aspects have been expanded and clarified in the revised manuscript through additional analyses and supporting figures and tables.

The uncertainty associated with the adopted precipitation forcing has been further evaluated. The imposed precipitation perturbations are defined relative to local mean precipitation conditions in order to ensure comparability across regions with different climatological backgrounds. To assess this approach, the relationship between the mean and standard deviation of monthly precipitation was analysed across all 199 grids. The results show that the mean and standard deviation of monthly precipitation are relatively highly consistent across the study area **(Fig. S3; Page 9, Lines 218 to 220)**, supporting the use of mean precipitation as a representative basis for defining the hypothetical precipitation perturbations. In addition, the use of pr_{rate} further reduces the direct influence of differences in background precipitation magnitude among regions.

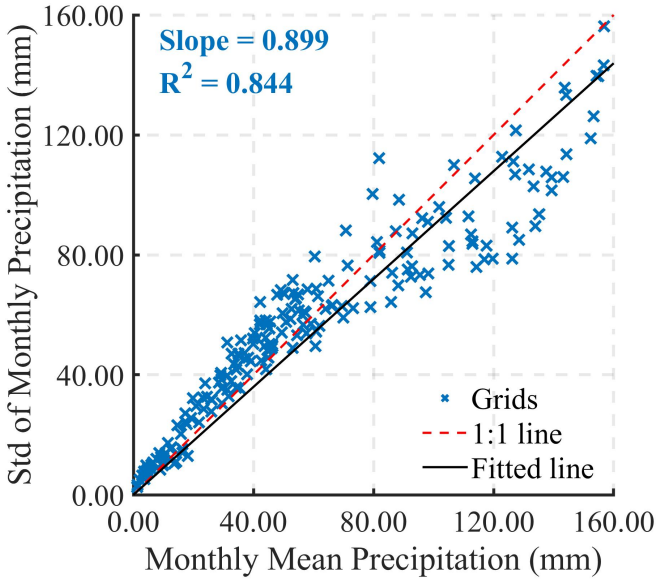


Figure S3: Relationship between mean and standard deviation of monthly precipitation across 199 grids. The red dashed line is the 1:1 line. The black solid line is the fitted line based on values from 199 grids. Blue crosses are the values for individual grids.

To address concerns regarding the significant of the differences, basin-scale Cohen's *d* analyses were further used to evaluate whether the differences among t_1~t_3 is attributable to random noise. The results show that, although differences among t_1~t_3 are small when responses are aggregated across the study area, they remain consistently distinguishable at the basin scale, with medium to large effect sizes across all river basins **(Fig. S4; Page 13, Lines 311 to 316)**. This indicates that these differences among t_1~t_3 cannot be attributed to random variability or noise.

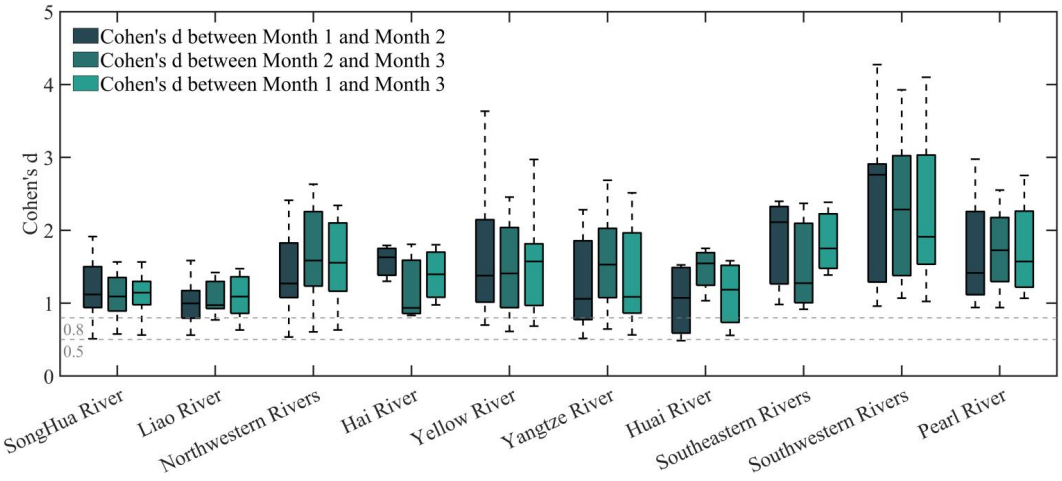


Figure S4: Basin-scale Cohen's d values for pairwise comparisons among early MYD forcing timings (t_1~t_3) across 10 major basins in China.

Regarding model accuracy, the basin-scale Cohen's *d* results indicate the differences among t_1~t_3 are distinguishable and cannot be attributed to random variability or noise **(Fig. S4; Page 13, Lines 311 to 316)**, supporting the existence of a separable early-stage response window. Based on this result, the robustness of the identified $t_{optimal}$ was evaluated using three independent approaches: the DAQM-derived estmates, the mode-based estimates, and the small perturbation-based estimates **(Fig. 5; Pages 14 to 15, Lines 322 to 362)**. Despite methodological differences, all approaches consistently indicate the dominance of t_1, while t_2 and t_3 remain non-negligible within the early-stage response window **(Fig. 5a; Fig. S5-S7; Pages 14 to 15, Lines 322 to 362)**. Comparison between the mode-based and DAQM-derived approaches shows that 75.88% of grids exhibit identical $t_{optimal}$ **(Fig. 5b; Pages 14 to 15, Lines 323 to 342).** The distribution of timing differences is approximately symmetric and centred around zero, indicating spatially random deviations without systematic bias **(Fig. 5b; Pages 23 to 24, Lines 519 to 523)**. Moreover, even when only grids with identical results are considered, t_2 and t_3 still account for a substantial proportion of $t_{optimal}$ (11.92% and 10.60%, respectively) **(Fig. S6)**. The comparison between small perturbation-based and DAQM-derived estimates further confirms robustness. Relative error in *k* remain generally small across China, with 72.86% of grids showing relative errors in *k* below 5% and most of the remaining grids around 6%, indicating relatively limited deviations overall. Although 27.14% of grids exhibit relative errors exceeding 5%, this does not affect $t_{optimal}$ identification, as 89.95% of grids show identical results between the two approaches **(Fig. 5c and 5d; Page 24, Lines 523 to 530)**. Overall, these results demonstrate strong robustness of the DAQM-derived $t_{optimal}$ across multiple independent evaluations. The DAQM provides a theoretical basis for interpreting and explaining the observed numerical patterns, offering a physically interpretable characterization of attenuation efficiency and $t_{optimal}$ identification.

Regarding the apparent asymmetry discussed for Fig. 7, we agree that the previous visualization may have exaggerated this perception because the original colorbar range was not symmetric around zero, which affected the visual interpretation of the magnitude differences and partially masked some grids with relatively larger positive differences. We recognize that this issue was overlooked in the original presentation. The figure has therefore been redrawn using a symmetric color scale in the revised manuscript **(now Fig. 6)**, allowing both positive and negative differences to be represented more

consistently and comparably. This revised visualization is also consistent with the probability distributions shown in **Fig. 7, Fig. S9 and Table. 1**. Importantly, based on the basin-scale Cohen's *d* results, pairwise differences among t_1~t_3 are distinguishable **(Fig. S4)**, indicating that the identified differences are not attributable to random variability or noise. Consistent with this, the spatial patterns shown in Fig. 5c and Fig. 5d demonstrate that these differences do not affect the identification of $t_{optimal}$, with 89.95% agreement between the small perturbation-based and DAQM-derived results, including in the north-western region **(Fig. 5c and 5d; Page 24, Lines 523 to 530)**. Furthermore, additional validation based on small perturbation experiments (pr_{rate} = 0.05) shows strong consistency with the full hypothetical precipitation-gradient framework, with correlation coefficients of R = 0.957 for Δk_{21} **(Fig. 6a and Fig. S8a)** and R = 0.955 for Δk_{31} **(Fig. 6b and Fig. S8b)** ($p<0.001$ in both cases) **(Page 16, Lines 379 to 381)**.

These additional analyses provide a more comprehensive basis for interpreting the magnitude, robustness, uncertainty, and spatial characteristics of the identified timing differences, and the corresponding clarifications have been incorporated throughout the revised manuscript.

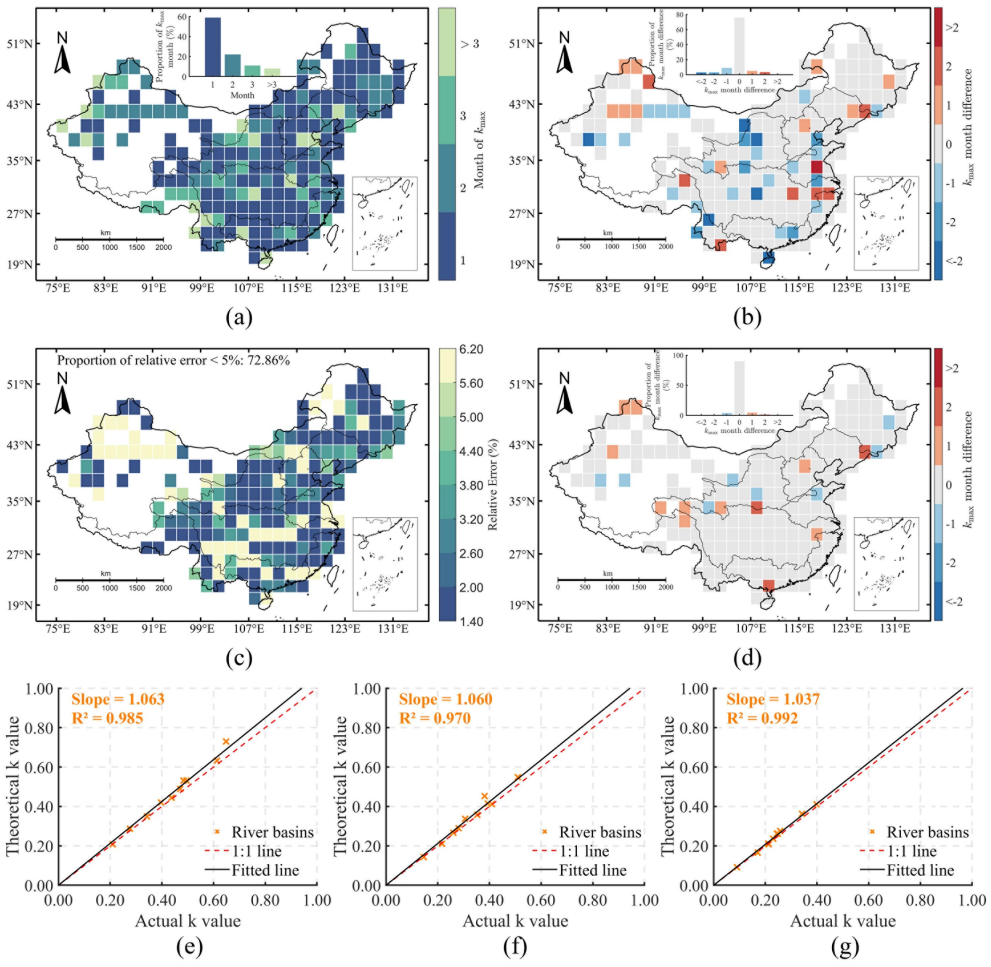


Figure 5: Validation and robustness assessment of the MYD attenuation efficiency coefficient (*k*). (a) Spatial distribution of the DAQM-derived $t_{optimal}$ for each grid. The inset shows the frequency histogram of the DAQM-derived $t_{optimal}$ across all grids. (b) Spatial distribution of $t_{optimal}$ differences (mode-based $t_{optimal}$ minus DAQM-derived $t_{optimal}$). The inset shows the frequency histogram of timing differences between

mode-based and DAQM-derived $t_{optimal}$ across all grids. (c) *Spatial distribution of* ε for each grid. (d) Spatial distribution of $t_{optimal}$ differences (small perturbation-based $t_{optimal}$ minus DAQM-derived $t_{optimal}$). The inset shows the frequency histogram of timing differences between small perturbation-based and DAQM-derived $t_{optimal}$ across all grids. (e-g) Median actual versus median theoretical k values for the ten major river basins in (e) t_1, (f) t_2, and (g) t_3. The red dashed line is the 1:1 line. The black solid line is the fitted line based on values from 10 river basins. Orange crosses are the values for individual river basins.

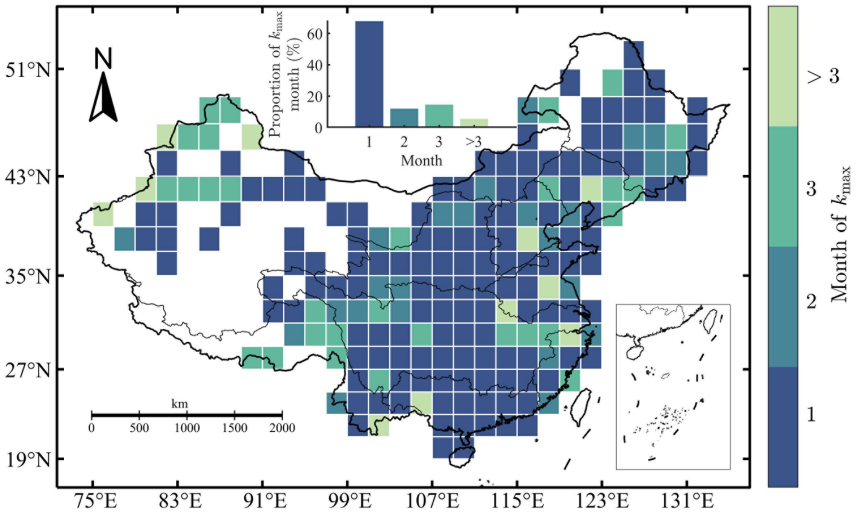


Figure. S5: Spatial distribution of the mode-based $t_{optimal}$ for each grid. For each hypothetical precipitation perturbation scenario (pr_{rate}= 0.05, 0.10, ..., 1.00), the corresponding optimal timing was first identified according to the maximum DS_{rate} among different $t_{forcing}$. Scenario without additional precipitation input (pr_{rate}= 0) was not included because identical DS_{rate} across all $t_{forcing}$ prevent a unique $t_{optimal}$ from being identified. The mode of the resulting 20 optimal timing was then used to represent the $t_{optimal}$ across different precipitation perturbation magnitudes (mode-based approach). The inset shows the frequency histogram of $t_{optimal}$ for these grids.

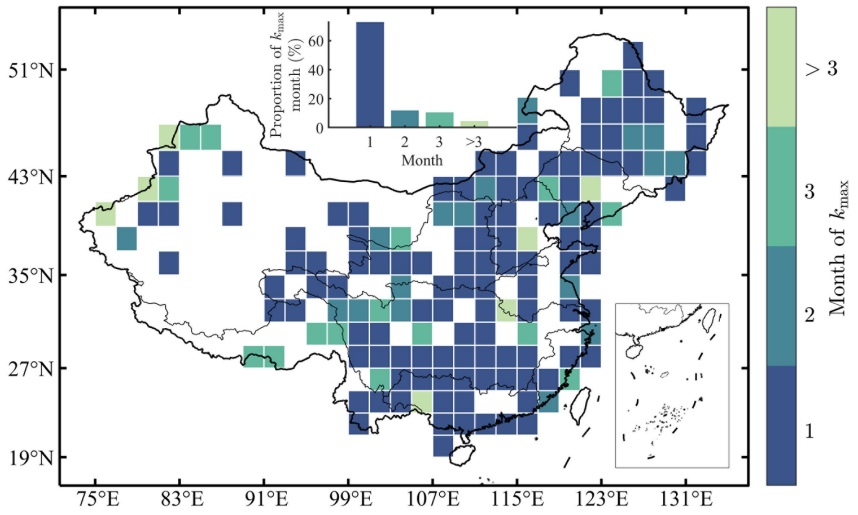


Figure. S6: Spatial distribution of the $t_{optimal}$ for grids where the mode-based and DAQM-derived $t_{optimal}$ are identical. The inset shows the frequency histogram of $t_{optimal}$ for these grids.

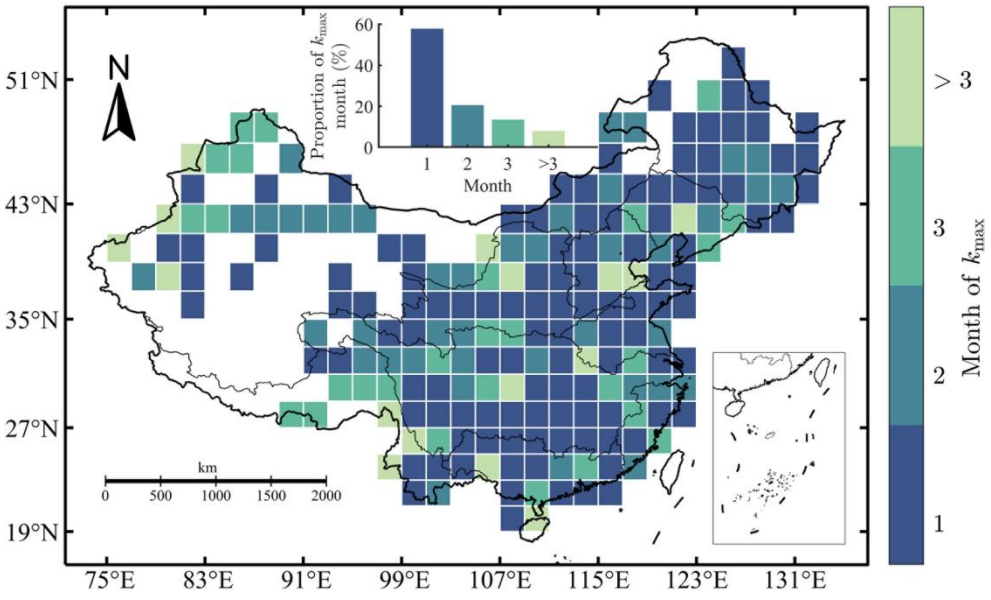


Figure. S7: Spatial distribution of the $t_{optimal}$ derived from small perturbation experiments (pr_{rate} = 0.05) for each grid. The inset shows the frequency histogram of $t_{optimal}$ for these grids.

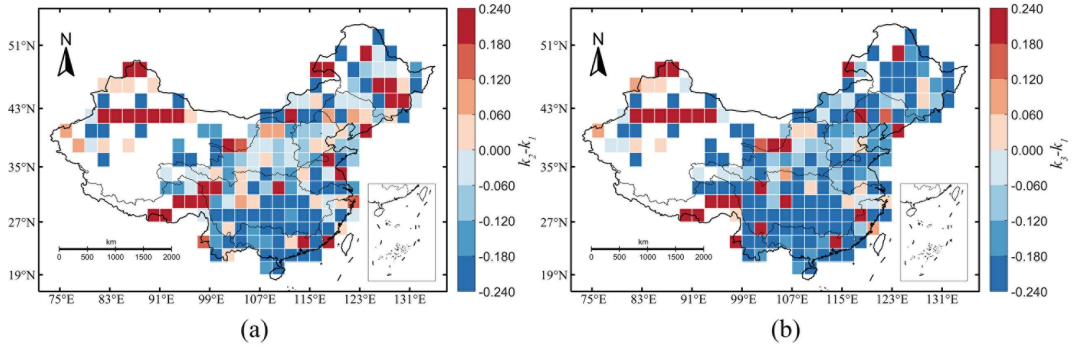


Figure 6: Spatial distribution of the changes in k (Δk) derived from the DAQM model. (a) Change in k between t_2 and t_1 (Δk_{21}). (b) Change in k between t_3 and t_1 (Δk_{31}).

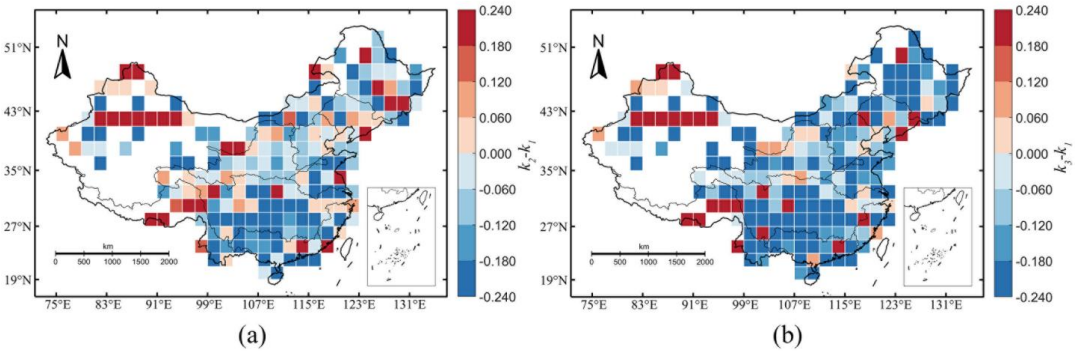


Figure. S8: Spatial distribution of the changes in k (Δk) derived from the small perturbation experiments (pr_{rate} = 0.05). (a) Change in k between t_2 and t_1 (Δk_{21}). (b) Change in k between t_3 and t_1 (Δk_{31}).

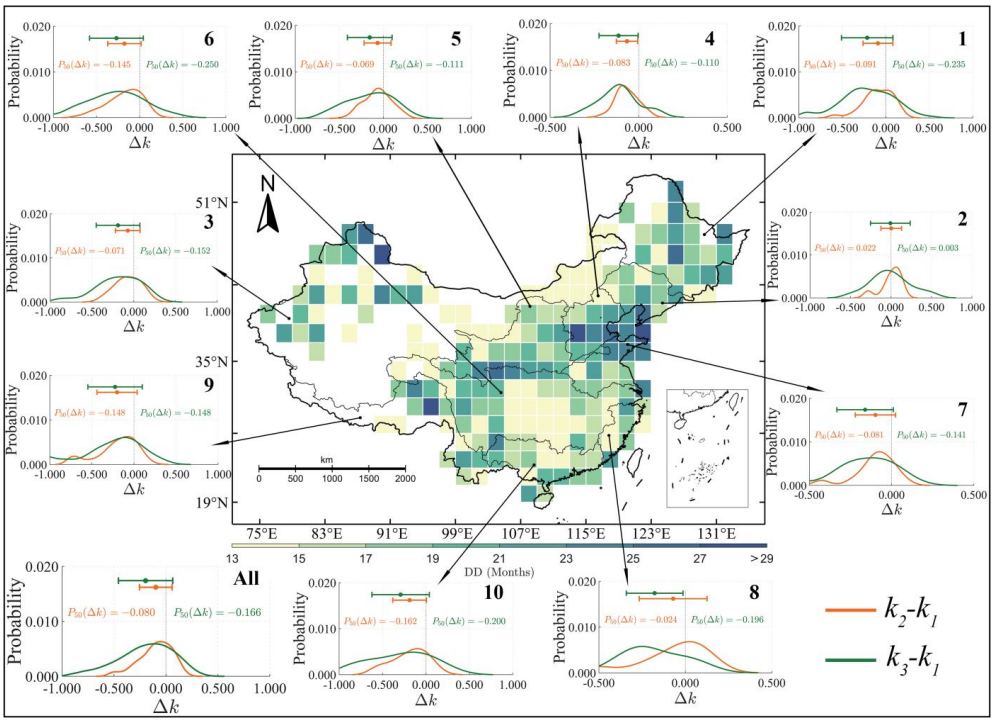


Figure 7: Spatial distribution of the median drought duration from 1991 to 2020 across the 10 river basins in China. Surrounding panels show the probability distribution of Δk derived from the DAQM model for all of China and each basin, based on grid-level MYD samples. Specifically, the orange lines represent Δk_{21}, and the green lines represent Δk_{31}. The *x-axis* in each panel denotes Δk, while the *y-axis* denotes probability. The number in the upper-right corner of each panel indicates the basin code.

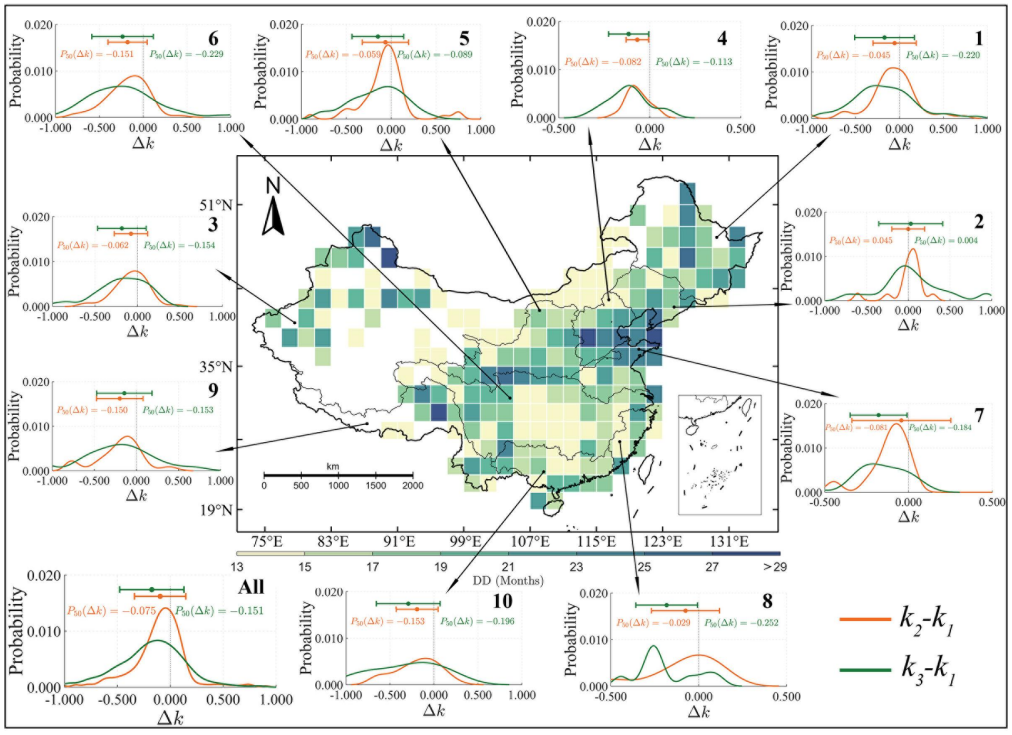


Figure. S9: Spatial distribution of the median drought duration from 1991 to 2020 across the 10 river basins in China. Surrounding panels show the probability distribution of Δk derived from the small perturbation experiments (pr_{rate} = 0.05) for all of China and each basin, based on grid-level MYD samples. Specifically, the orange lines represent Δk_{21}, and the green lines represent Δk_{31}. The *x-axis* in each panel denotes Δk, while the *y-axis* denotes probability. The number in the upper-right corner of each panel indicates the basin code.

Table 1: Summary of the number of grids, median values of DD, and changes in theoretical k (derived from the DAQM model; $\Delta k_{21}=k_2-k_1$, $\Delta k_{31}=k_3-k_1$, and $\Delta k_{32}=k_3-k_2$) and actual k (derived from the small perturbation experiments with pr_{rate} = 0.05), across the ten river basins and all of China.

River basin	Number of grids	Median DD (Months)	P_{50} (Δk_{21}) DAQM-derived	P_{50} (Δk_{21}) Small perturbation-based	P_{50} (Δk_{31}) DAQM-derived	P_{50} (Δk_{31}) Small perturbation-based	P_{50} (Δk_{32}) DAQM-derived	P_{50} (Δk_{32}) Small perturbation-based
Songhua River	26	22	-0.091	-0.045	-0.235	-0.220	***-0.142***	***-0.154***
Liao River	10	28	**0.022**	**0.045**	**0.003**	**0.004**	**-0.024**	**-0.026**
Northwestern Rivers	44	17	-0.071	-0.062	-0.152	-0.154	-0.087	-0.107
Hai River	8	29	-0.083	-0.082	-0.110	-0.113	-0.053	-0.054
Yellow River	20	21	-0.069	-0.059	-0.111	-0.089	-0.090	-0.081
Yangtze River	39	19	-0.145	-0.151	***-0.250***	-0.229	-0.127	-0.130
Huai River	11	23	-0.081	-0.081	-0.141	-0.184	-0.042	-0.039
Southeastern Rivers	8	14	-0.024	-0.029	-0.196	***-0.252***	-0.077	-0.108
Southwestern Rivers	15	21	-0.148	-0.150	-0.148	-0.153	-0.083	-0.085
Pearl River	18	19	***-0.162***	***-0.153***	-0.200	-0.196	-0.063	-0.077
All of China	199	21	-0.080	-0.075	-0.166	-0.151	-0.094	-0.093

Note: The bold value marks the maximum P_{50} among the ten river basins in China, and the bold-italic value marks the minimum P_{50}. P_{50} values are derived from both the DAQM-derived and small perturbation-based (pr_{rate} = 0.05) results.

Summary results in Fig. 5 mostly show almost identical results for t1-t3, and a huge jump to t4. This is never fully discussed. How is this justified? Can we reasonably analyse differences between t1-t3 within the uncertainty of such rough estimates? If the results between t1-t3 are barely distinguishable, what can we infer when one is selected over the other with minimal differences?

Response: We sincerely thank you for highlighting this important point regarding the interpretation of the relatively similar responses among t_1~t_3 and the much clearer separation from t_4 observed in Fig. 5. In the revised manuscript, this issue has been further clarified and discussed in greater detail.

In the original version, Fig. 5 was presented mainly using the mean responses fitted by Eq. 6 in order to illustrate the overall nonlinear attenuation behavior. However, both Eq. 6 and the subsequent analyses were originally developed based on drought periods at the grid scale (i.e., event-scale analysis), rather than solely on aggregated mean responses. We recognize that this aspect was not sufficiently reflected in the original presentation, which may have weakened the representation of differences among individual drought periods. In the revised manuscript, the figure and associated analyses have therefore been updated **(now Fig. 4a)** to more clearly emphasize the event-scale (grid-scale drought period) numerical experiment results.

Basin-scale Cohen's *d* analyses indicate that pairwise differences among t_1~t_3 are distinguishable and cannot be attributed to random variability or noise **(Fig. S4; Page 13, Lines 311 to 316)**. Since hypothetical precipitation inputs applied at a given $t_{forcing}$ can only influence the subsequent stage of MYD evolution, cumulative deficits were further calculated to quantify the drought impacts accumulated before each timing. Because PDSI values during MYD events remain below the drought threshold (PDSI<−1), cumulative absolute PDSI was used in the analysis. Based on this, the ratio of cumulative absolute PDSI from event onset to a given month within a single event relative to the DS of that event ($\frac{cumulative\ |PDSI|}{DS}$) was defined to quantify the proportion of drought impacts accumulated before each $t_{forcing}$. The ratio increased progressively with MYD development across all MYD events **(Fig. 4b)**. The $\frac{cumulative\ |PDSI|}{DS}$ values corresponding to t_1, t_2, t_3, and t_4 were approximately 0.04, 0.10, 0.17, 0.25, respectively **(Page 13, Lines 302 to 310)**. These results indicate that a progressively larger fraction of DS had already accumulated by later $t_{forcing}$ during MYD development, thereby reducing the remaining alleviation potential. Taken together, these results support the existence of an early-stage response window (t_1~t_3) during MYD evolution **(Page 22, Lines 482 to 487)**.

Within this early-stage response window (t_1~t_3), the relatively close but still distinguishable responses reflect the competing influence of cumulative temporal advantage associated with earlier precipitation input (Conceptual Model 1) and the higher numerical sensitivity of standardized drought indices under more severe drought conditions (Conceptual Model 2), as discussed in the revised manuscript **(Page 22,**

Lines 462 to 477). Progressively increasing $\frac{cumulative\ |PDSI|}{DS}$ during drought development (e.g., approximately 0.04, 0.10, 0.17, and 0.25 at t_1, t_2, t_3, and t_4, respectively) indicate the continuous accumulation of drought impacts during MYD evolution. Once $\frac{cumulative\ |PDSI|}{DS}$ reaches a sufficiently high level (approximately 0.25 at t_4), the cumulative temporal advantage associated with earlier precipitation input becomes dominant, leading to a clearer separation among $t_{forcing}$ **(Page 22, Lines 482 to 487)**.

Overall, our results do not suggest that "later is better" for MYD alleviation. Instead, they reveal the existence of an early-stage response window during which t_2 and t_3 may still exhibit attenuation efficiencies comparable to or even exceeding that of t_1. This finding is particularly important for MYDs because these events involve cumulative water deficits across seasons and years, making the timing of precipitation input especially critical for drought alleviation. Missing t_1 does not necessarily imply that the opportunity for maximum attenuation efficiency has already been lost, provided that precipitation input still occurs within the early-stage window **(Page 23, Lines 506 to 511)**.

These clarifications have now been incorporated into the revised manuscript.

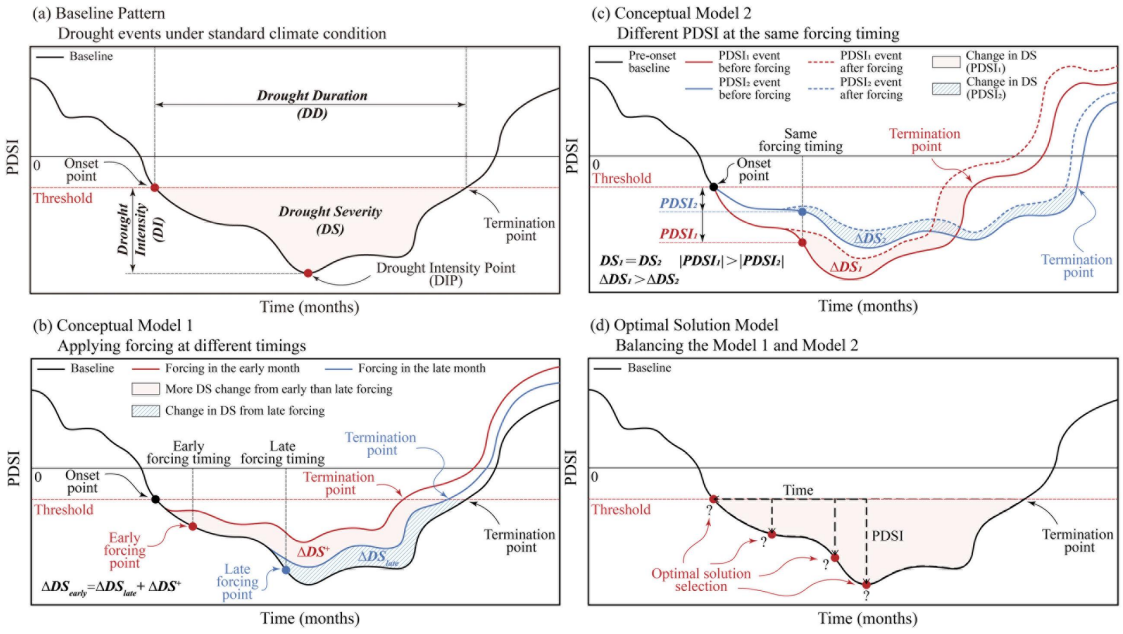


Figure S2: Conceptual models illustrating MYD alleviation driven by hypothetical increased precipitation forcing: (a) Baseline pattern without increased forcing. *DS* is represented by the area of the red region. (b) Conceptual model 1: The same hypothetical precipitation increase applied at different $t_{forcing}$ within an MYD event. An earlier $t_{forcing}$ leads to greater alleviation benefits (ΔDS^+, red area) and avoids additional losses from cumulative precipitation deficits. (c) Conceptual model 2: The same hypothetical precipitation increase applied at the same $t_{forcing}$, but with different drought index values. Because standardized drought indices exhibit tail sensitivity, deeper drought conditions tend to produce large DS reductions under identical forcing ($\Delta DS_1 > \Delta DS_2$). (d) Optimal solution model: By weighing the mismatch influences of hypothetical precipitation forcing described in Conceptual models 1 and 2, $t_{optimal}$ (marked by the red point) can be identified.

In its current status, I do not see major value in the proposed scientific outcomes, and only a substantial revision of the content of the manuscript, with a proper quantification of the uncertainty in the produced outcomes, may justify a publication in my opinion.

Response: We sincerely thank you for this comment and have carefully addressed the concerns regarding scientific value of the study and the quantification of uncertainty. In the revised manuscript, substantial modifications have been made through additional uncertainty analyses, robustness evaluations, and related revisions.

Specially, the revised manuscript now includes statistical significance testing **(Page 12, Lines 289 to 298)**, effect size analysis **(Pages 12 to 13, Lines 289 to 316)**, and comparisons among DAQM-derived, mode-based, and small perturbation-based $t_{optimal}$ identification **(Pages 14 to 15, Lines 322 to 362; Pages 23 to 24, Lines 512 to 535)**. These analyses consistently show that, although the differences among t_1~t_3 remain relatively small when aggregated across all MYD events over the study area, they are still distinguishable at the river basin scale and cannot be attributed to random variability or noise. The results also show relatively high consistency among multiple independent estimates of $t_{optimal}$, supporting the robustness of the identified early-stage response window.

The revised manuscript also explicitly discuss the competition between the cumulative temporal advantage associated with earlier precipitation input and the higher numerical sensitivity of standardized drought indices under more severe drought conditions, which help explain the existence of the early-stage response window among t_1~t_3 **(Page 22, Lines 462 to 484)**.

Accordingly, the revised manuscript does not suggest that "later is better" for MYD alleviation, nor does it aim to identify a universally optimal month under all conditions. Instead, the results indicate the existence of an early-stage response window during which t_2 and t_3 may still exhibit attenuation efficiencies comparable to, or occasionally exceeding, that of t_1. These findings provide a quantitative basis for understanding how drought attenuation efficiency evolves during MYD development.

Specific comments

L82-83. The work of Zhang et al. (2024) is unproperly referred, as they discuss about effects in difference climate regions and not "wet and dry extreme conditions".

Response: We sincerely thank you for pointing this out. The citation to Zhang et al. (2024) has been removed in the revised manuscript to avoid potential inconsistency with the original scope of the cited study.

L199-214. The role of these two conceptual models is unclear. As the results are based on the numerical experiments and the exponential model, this discussion early in the manuscript is confusing. My understanding is that you are using these two models to justify the results that, sometimes, t2 and t3 are more optimal than t1. However, this should be reported and discussed later, and only when a significant difference in the results is observed (see general comment). By reporting this discussion here, a reader would expect two different modelling approaches, which is not the case for your study.

Response: We sincerely thank you for this insightful comment. In the revised manuscript, the discussion of the two conceptual models has been moved from the methodological description to the Discussion section and is now presented after the corresponding numerical experiment results and uncertainty analyses **(Pages 12 to 16, Lines 277 to 372)** has been introduced **(Page 22, Lines 462 to 477)**. This revision clarifies that the conceptual models are not independent modelling approaches, but are instead used to interpret the early-stage response window identified from the numerical experiments.

L230. This sentence is very confusing. My understanding is that "baseline" and "forcing" in eqs. (2-7) refer to the actual conditions and the one obtained by altering the precipitation forcing. If this is the case, it is not clear to me why the baseline period 1961-1990 cannot be used as well. If, instead, the "baseline" refers to 1961-1990 and the "forcing" to 1991-2020 (with altered precipitation), this will introduce a series of issue related to non-stationarity, inconsistency in the two datasets, etc. You should clarify better how the data are analysed, and why the period 1961-1990 is excluded.

Response: We sincerely thank you for this important comment. The definitions of the calibration period, baseline scenario, and forcing scenario have been clarified in the revised manuscript to avoid ambiguity.

In the calculation of the Palmer Drought Severity Index, the climatologically appropriate for existing conditions (CAFEC) precipitation is determined based on a selected calibration period and is used to derive the weighting factors in the Z-index. Both the use of a full-record calibration period and a fixed historical calibration period are acceptable approaches in PDSI calculation. In this study, a fixed calibration period (1961-1990) is adopted to ensure consistency of these coefficients across all numerical experiment scenarios, consistent with commonly adopted choices such as NOAA's use (Karl, 1986) **(Page 7, Lines 162 to 163)**. This calibration period is not used as the baseline scenario in the numerical experiments.

The baseline and forcing scenarios used in the numerical experiment are both derived from the period 1991-2020. Precipitation perturbations are applied within this period to construct the forcing scenarios, while the baseline corresponds to the unperturbed conditions over the same period. Using the full record to calculate CAFEC would cause the calibration itself to be influenced by the imposed precipitation perturbations, thereby preventing a consistent reference state. The use of a fixed calibration period avoids this issue and ensures that all scenarios are evaluated against the same climatological reference, as illustrated in **Fig. S1** in the Supplementary Material **(Page 7, Lines 163 to 164)**.

These definitions and their roles have been clarified in the revised manuscript to avoid ambiguity.

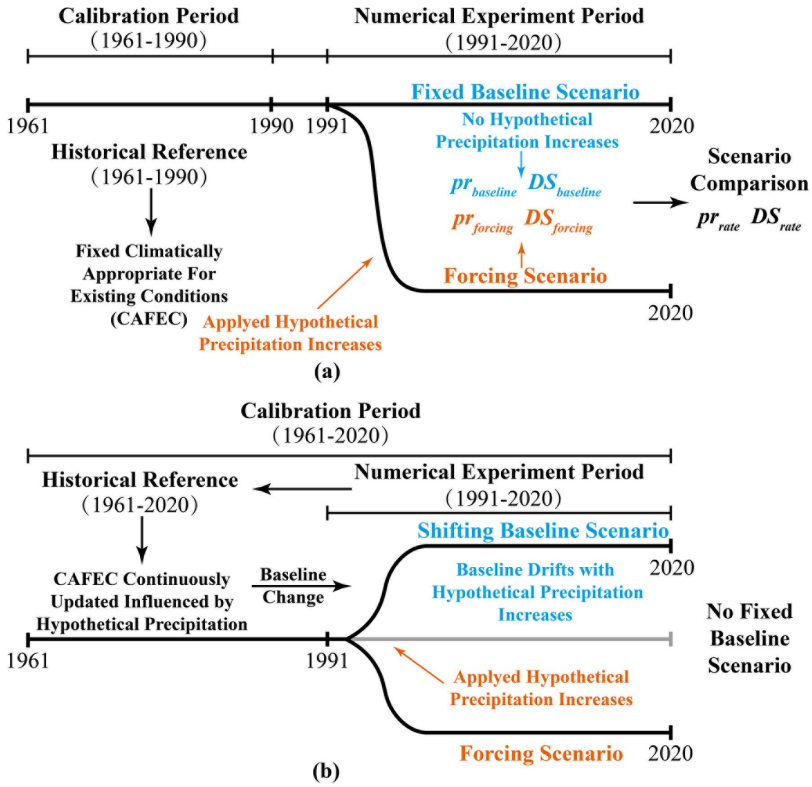


Figure S1: Conceptual illustration of the experimental design. Panel (a) shows a fixed baseline derived from a constant calibration period, while panel (b) illustrates a shifting baseline that arises when the calibration is affected by hypothetical precipitation increases imposed in the numerical experiment.

Karl, T. R.: The Sensitivity of the Palmer Drought Severity Index and Palmer's Z-Index to their Calibration Coefficients Including Potential Evapotranspiration, J. Appl. Meteorol., 25, 77–86, doi: https://doi.org/10.1175/1520-0450(1986)025<0077:TSOTPD>2.0.CO;2, 1986.

Eq. (2). Some unrealistic precipitation values can be simulated with this equation, as up-to 100% of the average is added to the actual value, which can reach a final value close to 200% of the average (when the base is just slightly lower than average). Standard deviation should be used for this purpose instead.

Response: We sincerely thank you for this valuable comment. The formulation in Eq. (2) is designed to represent a controlled range of hypothetical precipitation perturbations within the numerical experiment framework. Considering that MYD responses may vary under different precipitation perturbation magnitudes, pr_{rate} ranging from 0 to 100% at 5% increments were used to systematically evaluate MYD responses across different perturbation conditions, thereby integrating both small and relatively large hypothetical precipitation perturbations. While the upper bound may lead to values approaching twice the mean precipitation under certain conditions, such magnitudes remain physically plausible within the context of extreme precipitation variability. Extreme precipitation events can substantially exceed mean climatological conditions in observational records, indicating that the upper bound of the imposed perturbations remains within a physically plausible range of variability (Nie and Sun, 2022; Gong et al., 2025; Hou et al., 2025).

The imposed precipitation perturbations are defined relative to local mean precipitation conditions in order to ensure comparability across regions with different climatological backgrounds. We agree that standard deviation is also an important measure of precipitation variability. To further evaluate this issue, the relationship between mean and standard deviation of monthly precipitation was analysed across all 199 grids. The results show that the mean and standard deviation of monthly precipitation are relatively highly consistent across the study area **(Fig. S3)**, supporting the use of mean precipitation as a representative and comparable basis for defining the hypothetical precipitation perturbations.

These considerations and corresponding supporting analyses have been clarified in the revised manuscript **(Page 9, Lines 218 to 223)**.

Gong, Y., Fan, K., and Yang, H.: The causes of the extreme precipitation event in the Sichuan Basin in August 2020: A perspective from anomalous inter-monthly climate differences and weather processes, Atmos. Res., 108450. doi: https://doi.org/10.1016/j.atmosres.2025.108450, 2025.

Hou, R., Zhang, W., Hu, S. and Xu, R.: Rapid Indian Ocean warming fuels more frequent extreme pre-flood season rainfall over southern China, Natl. Sci. Rev., 12, nwaf298. doi: https://doi.org/10.1093/nsr/nwaf298, 2025.

Nie, Y. and Sun, J.: Moisture sources and transport for extreme precipitation over Henan in July, Geophys. Res. Lett., 49, e2021GL097446. doi: https://doi.org/10.1029/2021GL097446, 2022.

Eq. (3). DS_rate and Pr_rate are defined somewhat inconsistently, as the second is Pr_rate = (Pr_base-Pr_forcing)/Pr_mean according to eq. 2. The two rates should be consistently defined.

Response: We sincerely thank you for this comment. The definition of DS_{rate} and pr_{rate} have been revised to ensure consistency. The corresponding equations and descriptions have been updated accordingly in the revised manuscript **(Page 9, Lines 215 to 218)**.

In this study, the formulation of DS_{rate} was arranged such that drought attenuation is represented as a positive quantity. For drought events, increased hypothetical precipitation leads to reduced DS, meaning that $DS_{forcing}$ is smaller than $DS_{baseline}$. This treatment does not affect the correspondence between pr_{rate} and DS_{rate}.

L294. It is not clear how eq. 6 is applied. Is it applied for each grid cell using all the drought events and scenarios for a given t_forcing? For each event separately? Why the data in fig. 5 report the theoretical average and not the empirical values (or both)? To improve the clarity of your methodology, results for two grid cells should be reported, showing the empirical and theoretical values for different t, and contrasting a case where t1 is optimal against one where t3 is optimal. This would help supporting the statement that the differences are relevant/significant.

Response: We sincerely thank you for this comment. Eq. 6 is applied at the event level, where for each drought period and for each forcing time ($t_{forcing}$), the relationship is evaluated across all precipitation perturbation scenarios **(Page 12, Lines 283 to 284)**.

In the original version, Fig. 5 was presented mainly using the mean responses fitted by Eq. 6 in order to illustrate the overall nonlinear attenuation behavior. However, both Eq. 6 and the subsequent analyses were originally developed based on drought periods at the grid scale (i.e., event-scale analysis), rather than solely on aggregated mean responses. We recognize that this aspect was not sufficiently reflected in the original presentation, which may have weakened the representation of differences among individual drought periods. In the revised manuscript, the figure and associated analyses have therefore been updated **(now Fig. 4a)** to more clearly emphasize the event-scale (grid-scale drought period) numerical experiment results. Fig. 4a now presents only the empirical responses derived directly from the numerical experiments. This revision was made to avoid the impression that the DAQM was fitted directly to aggregated mean responses across the study area.

L308-318. It is not clear to me what is the added value of this analysis per "stages" compared to what has been already said. The stages seem to be defined somewhat arbitrarily, and the decreasing added value was already discussed in the previous paragraph. Maybe I am missing something, but a rewording may be needed.

Response: We sincerely thank you for this comment. This section has been revised in the updated manuscript. In the original version, the role of the stage-based analysis was not sufficiently clear and may have appeared repetitive relative to the previous discussion of decreasing drought alleviation efficiency.

In the revised manuscript, the analysis has been restructured to better clarify its role in the explaining the observed attenuation patterns **(Pages 12 to 13, Lines 281 to 316)**. The revised analysis now focuses on quantifying the cumulative impacts accumulated before different $t_{forcing}$ and evaluating whether the relatively close responses among t_1~t_3 remain distinguishable at the basin scale. These revisions were intended to better explain the emergency of the early-stage response window and the clearer reduction in attenuation efficiency from t_4 onward.

Fig. 5. There is a clear difference between t1-t3 and t4. Why is that? What is so different between rains at t3 and t4 that justify such result? Is this a general result or specific of your dataset? No discussion on this behaviour is reported in the current version of the manuscript.

Response: We sincerely thank you for highlighting this important point. This behavior has been explicitly clarified and discussed in the revised manuscript.

Basin-scale Cohen's *d* analyses indicate that the differences among t_1~t_3 are distinguishable and cannot be attributed to random variability or noise **(Fig. S4; Page 13, Lines 311 to 316)**. In addition, the use of pr_{rate} further improves the comparability of attenuation responses among different timing within different MYD events across different river basins.

Mechanistically, the relatively close responses among t_1~t_3 reflect the competition between the cumulative temporal advantage associated with earlier precipitation input (as described in Conceptual

Model 1) and the higher numerical sensitivity of standardized drought indices under more severe drought conditions (as described in Conceptual Model 2). During the first three months, the effects described in Conceptual Models 1 and 2 partially offset each other, producing relatively close but still distinguishable attenuation response. However, progressively increasing $\frac{cumulative\ |PDSI|}{DS}$ during drought development (e.g., approximately 0.04, 0.10, 0.17, and 0.25 at t_1, t_2, t_3, and t_4, respectively) indicate the continuous accumulation of drought impacts during MYD evolution. Once $\frac{cumulative\ |PDSI|}{DS}$ reaches a sufficiently high level (approximately 0.25 at t_4), the cumulative temporal advantage associated with earlier precipitation input becomes dominant, leading to the clearer reduction in attenuation efficiency observed after t_1~t_3. The consistency among the DAQM-derived results, the mode-based results, and the small perturbation-based results suggests that the observed MYD response patterns are not dependent on a single estimation approach, but remain robust across independent evaluations **(Page 13, Lines 302 to 310; Pages 22 to 23, Lines 459 to 488)**.

The corresponding discussion has been added to the revised manuscript to clarify this behavior.

L330-334. Is there any grid cells that show a difference in t_optimal using the empirical k instead of the theoretical values? How many? Are the differences significantly large (e.g., comparable with the differences t1-t2/t3)?

Response: We sincerely thank you for this insightful comment. To clarify, a limited number of grids exhibit differences in $t_{optimal}$ when comparing the DAQM-derived, mode-based, and small perturbation-based results.

The basin-scale Cohen's *d* results indicate the differences among t_1~t_3 are distinguishable and cannot be attributed to random variability or noise **(Fig. S4; Page 13, Lines 311 to 316)**, supporting the existence of a separable early-stage response window. Based on this result, the robustness of the identified $t_{optimal}$ was evaluated using three independent approaches: the DAQM-derived estmates, the mode-based estimates, and the small perturbation-based estimates **(Fig. 5; Pages 14 to 15, Lines 322 to 362)**. Despite methodological differences, all approaches consistently indicate the dominance of t_1, while t_2 and t_3 remain non-negligible within the early-stage response window **(Fig. 5a; Fig. S5-S7; Pages 14 to 15, Lines 322 to 362)**. Comparison between the mode-based and DAQM-derived approaches shows that 75.88% of grids exhibit identical $t_{optimal}$ **(Fig. 5b; Pages 14 to 15, Lines 323 to 342).** The distribution of timing differences is approximately symmetric and centred around zero, indicating spatially random deviations without systematic bias **(Fig. 5b; Pages 23 to 24, Lines 519 to 523).** Moreover, even when only grids with identical results are considered, t_2 and t_3 still account for a substantial proportion of $t_{optimal}$ (11.92% and 10.60%, respectively) **(Fig. S6)**. The comparison between small perturbation-based and DAQM-derived estimates further confirms robustness. Relative error in *k* remain generally small across China, with 72.86% of grids showing relative errors in *k* below 5% and most of the remaining grids around 6%, indicating relatively limited deviations overall. Although 27.14% of grids exhibit relative errors exceeding 5%, this does not affect $t_{optimal}$ identification, as 89.95% of grids show identical results between the two approaches **(Fig. 5c and 5d; Page 24, Lines 523 to 530)**. Furthermore, additional validation based on small perturbation experiments (pr_{rate} = 0.05)

shows strong consistency with the full hypothetical precipitation-gradient framework, with correlation coefficients of R = 0.957 for Δk_{21} **(Fig. 6a and Fig. S8a)** and R = 0.955 for Δk_{31} **(Fig. 6b and Fig. S8b)** ($p<0.001$ in both cases) **(Page 16, Lines 379 to 381)**. Overall, these results demonstrate strong robustness of the DAQM-derived $t_{optimal}$ across multiple independent evaluations. The DAQM provides a theoretical basis for interpreting and explaining the observed numerical patterns, offering a physically interpretable characterization of attenuation efficiency and $t_{optimal}$ identification.

The corresponding discussion has been added to the revised manuscript to further clarify this point.

A final comment is on the general question "is early always better?". Even if the results would be able to show - without the previously mentioned doubts related to uncertainty, methodology, etc. – that t2 and t3 are sometime better than t1, would this be enough to support that early is not always better? The focus here is on multi-year drought events, e.g., events lasting mor than 12 months, sometime multiple years. In this framework, aren't the first 3 months still early? No considerations are reported on the practical implications of an optimal at t2 compared to t1, which is an odd omission in a manuscript with such title.

Response: We sincerely thank you for this important comment. In the context of MYD events, the first several months represent an early-stage response window (t_1~t_3). Our results do not suggest that "later is better" for MYD alleviation. Instead, they reveal the existence of an early-stage response window during which t_2 and t_3 may still exhibit attenuation efficiencies comparable to or even exceeding that of t1. Missing t_1 does not necessarily imply that the opportunity for maximum attenuation efficiency has already been lost, provided that precipitation input still occurs within the early-stage window. This finding is particularly important for MYDs because these events involve cumulative water deficits across seasons and years, making the timing of precipitation input especially critical for drought alleviation. The $\frac{cumulative\ |PDSI|}{DS}$ values corresponding to t_1, t_2, t_3, and t_4 (approximately 0.04, 0.10, 0.17, 0.25, respectively) further indicate that meaningful drought accumulation has already occurred within the early-stage window. Once this accumulation reaches a higher level (around 0.25), the response transitions beyond this early-stage response window, resulting in substantially reduced alleviation potential. MYD alleviation therefore exhibits a flexible early-stage response window during which attenuation efficiency remains relatively high within t_1~t_3. These findings provide a more nuanced understanding of drought alleviation timing, highlighting both the importance of early-stage responses and the flexibility that reflects the complexity of drought processes.

The corresponding discussion has been added to the revised manuscript **(Pages 22 to 23, Lines 459 to 511)**.