REVIEWER COMMENTS

This manuscript provides valuable and much-needed insights into the sub daily variability and the diurnal cycle of precipitation across satellite and reanalysis estimates. I particularly appreciate that the authors go beyond routine intercomparisons by leveraging physical constraints to uncover the mechanisms driving observed product strengths and limitations. However, to help the manuscript reach its full potential, I would encourage the authors to refine its overall narrative structure. Currently, the study attempts to balance three distinct analytical angles simultaneously, ground validation over Northern Hemisphere land domains (CONUS, Europe, East Asia), global-scale unvalidated intercomparisons across land and ocean, and process-oriented physical diagnostics at two U.S. ground sites. Combining these different scopes creates some narrative fragmentation. In particular, the global ocean/land intercomparison adds limited new insight, given similar existing literature and strays slightly from the core focus nicely anchored by the paper's title ("Evaluation against ground-based observations"). Streamlining these components would reduce visual clutter and allow the true highlight of this work moving beyond descriptive bias mapping to explain the physical drivers of diurnal discrepancies to shine through much more clearly.

➔ We thank the reviewer for identifying the lack of a clear connection among the different parts of the original manuscript. We agree that the ground-based evaluation, the broader inter-product comparison, and the process analysis were previously presented as separate components rather than as a continuous argument. In revising the manuscript, we made the comparison with ground-based observations, the main basis of the results. In addition to the spatially averaged statistics, the revised figures now show the difference of the Multi-Satellite and Multi-Model means against the ground-based observations at station-collocated grid cells. This makes the magnitude and regional structure of the dataset biases directly visible.

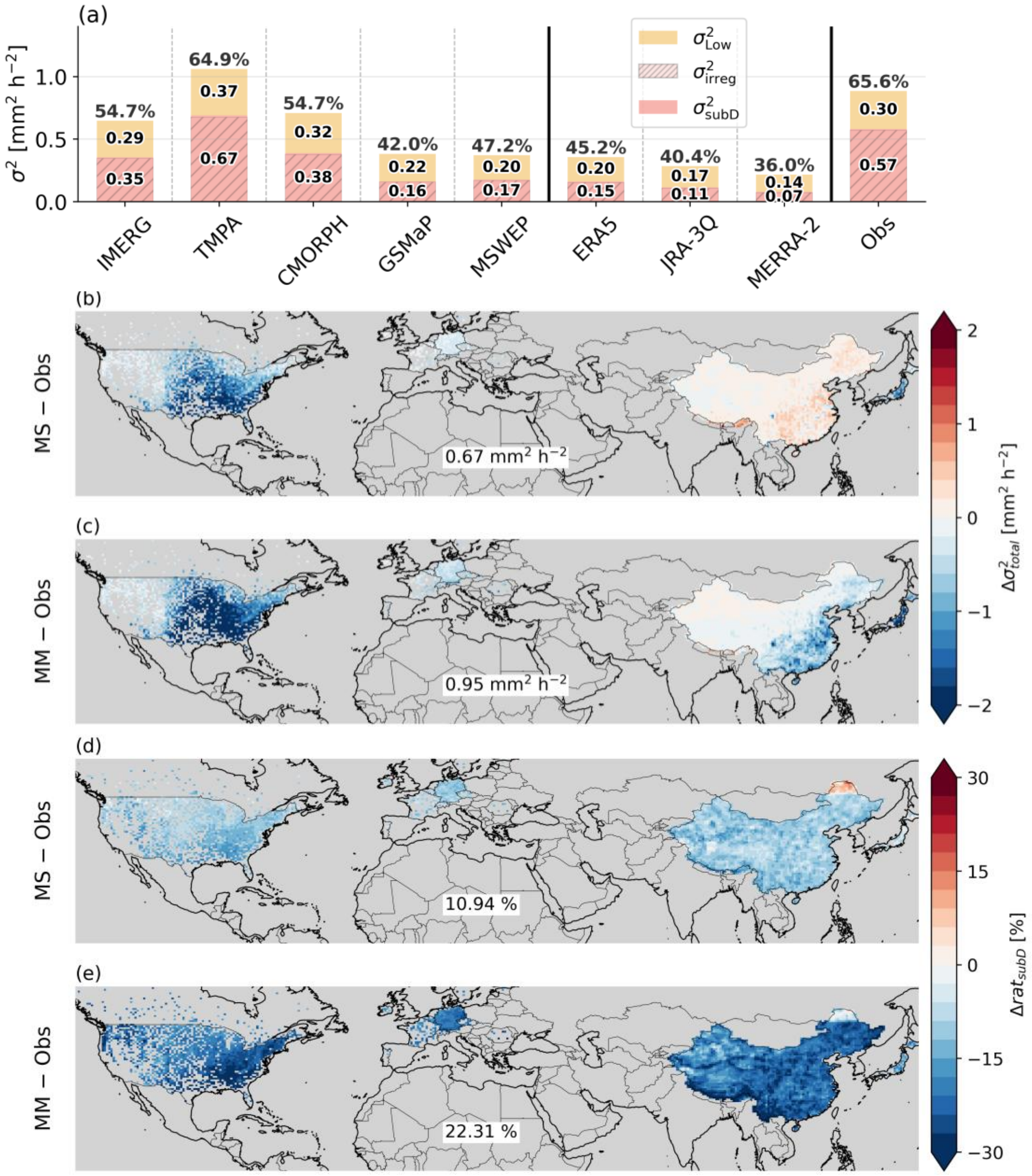
The subsequent analyses have been reorganized to explain the discrepancies identified from this comparison. Individual dataset maps and other supporting results have been moved to the Supplement, while the main text focuses on the Multi-Satellite and Multi-Model comparisons and their differences from the observations. The precipitation-type decomposition then identifies whether the phase differences originate from convective or large-scale precipitation, and the thermodynamic and moisture-budget analyses examine why these differences occur under contrasting precipitation regimes. The revised Results section therefore proceeds from direct evaluation against observations to the physical interpretation of the identified biases.

We retained the quasi-global inter-product analysis only as a complementary part of the study. Its purpose is to determine whether the characteristics identified over the observed land regions are also present over areas where comparable ground-based observations are unavailable. We have made this distinction explicit in the Abstract and Introduction and do not present the quasi-global comparison as an observation-based evaluation. The detailed changes are described in our responses to the specific comments below.

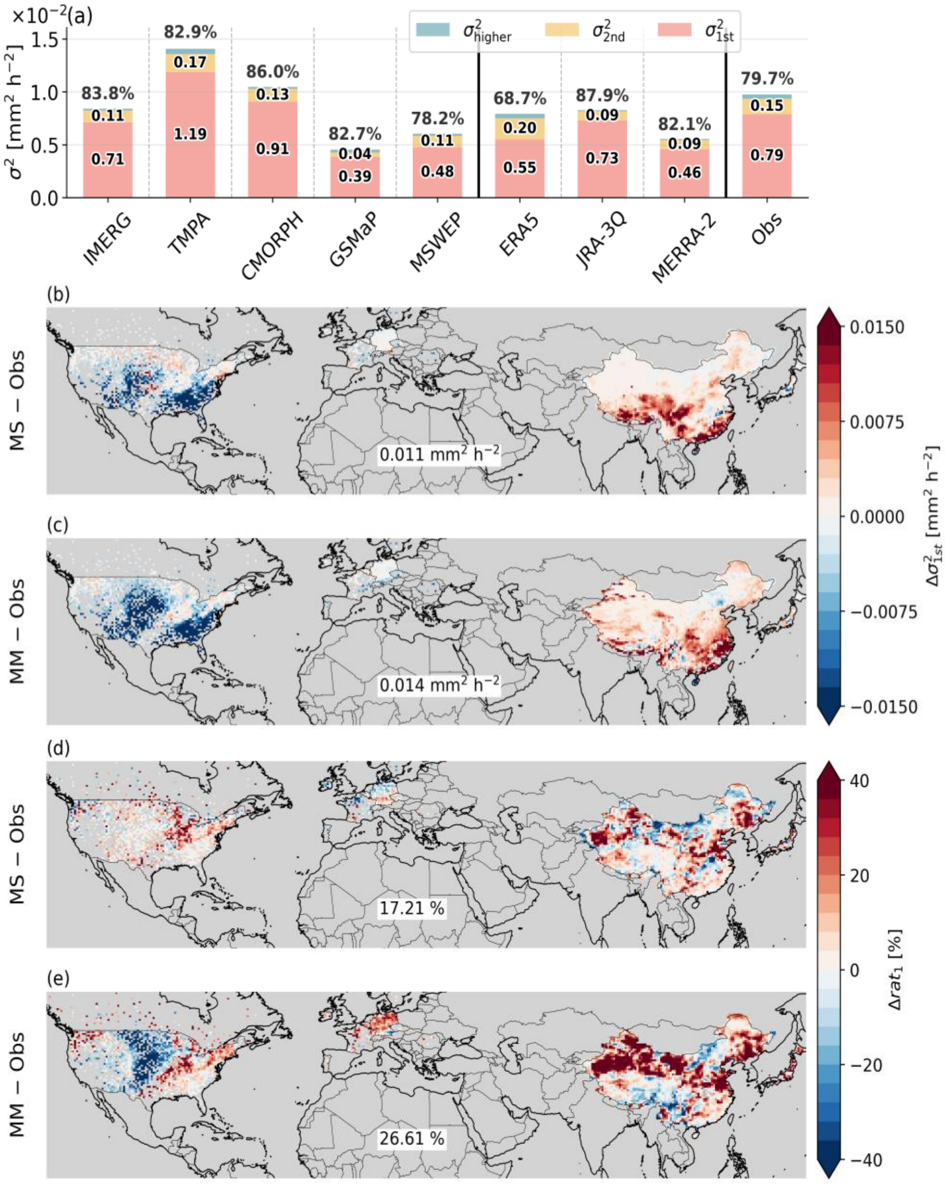
Major comments

1. Given that the title and scope of the manuscript emphasize comparison with ground observations, I expected a direct spatial comparison map against the gauge network (Model − Obs and Satellite − Obs). Providing only spatially averaged bar plots (e.g., Fig. 3a, 4a) masks all regional performance details.

➔ We agree with reviewer's comment, so in this revision, we try to compare the gridded precipitation datasets to the collocated ground-based observations of every grid cell containing at least one station. The spatial maps of their differences now appear in the main text alongside the bar plots they accompany. Figure 4 retains the bar statistics in panel (a) and added the MS and MM means minus the observations for the total variance σ^2_{total} (panels b, c) and for the sub-daily variance ration rat_{subD} (panels d, e). Figure 6 is the same for the 1st harmonic variance σ^2_{1st} (panels b, c) and for rat_1 (panels d, e). Each panel is annotated with the spatial RMSE over the grid cells shown. In addition, we also include the spatial diurnal phase difference of the MS and MM means to the ground observations in Figs. 9c and 9d.



"Figure 4: (a) Sub-daily (pink) and lower-frequency (orange) variance components of precipitation at observation locations. Sub-daily frequency variance is obtained via high-pass Butterworth filter. Numbers inside bars indicate variance magnitude mm² h⁻², and percentage above bars represent the sub-daily variance ratio (rat_{subD}). Differences from the ground-based observations, for (b, d) the Multi-Satellite (MS; IMERG, TMPA, CMORPH, GSMaP, MSWEP) and (c, e) the Multi-Model (MM; ERA5, JRA-3Q, MERRA-2) mean, in terms of (b, c) total variance (σ^2_{total}) and (d, e) rat_{subD}, where the spatial averaged value is noted in each panel below."



"Figure 6: (a) Same as Fig.4a, but for regular sub-daily precipitation variance (σ^2_{reg}) decomposed into the 1st (pink), 2nd (orange), and higher-order (blue) harmonic components. (b–e) Same as Fig. 4b–e, but for (b, c) the 1st harmonic variance (σ^2_{1st}) and (d, e) its fractional contribution to σ^2_{reg} (rat_1)."

➔ The maps show that the two quantities behave differently. For rat_{subD} the difference is negative at nearly every station-collocated cell, by 10.3% for the MS and 21.9% for the MM, and the sign does not change across all researched areas, so the spatially average bars are representative of the region covered by the stations. The magnitude of the MM underestimation also holds at 20–30% with little regional variation, in contrast to σ^2_{total}, whose underestimation peaks over the central CONUS. For the spatial difference in σ^2_{1st}, central and eastern CONUS exhibit the underestimated σ^2_{1st} in the MS and MM and its overestimation is displayed in southern China. On the other hand, the MM represents a general overestimation in the validation areas, except for the underestimation over the GP region, but such discrepancy is highly diminished in the MS. The corresponding description is updated in main text as follows:

Lines 534–546: ***"The spatially aggregated statistics in Fig. 4a do not indicate where these discrepancies arise, so the Multi-Satellite (MS) and Multi-Model (MM) means are differenced against the ground-based observations over the regions of dense gauge coverage (Figs. 4b–e). Both product groups underestimate σ^2_{total} over most of CONUS, and the deficit is largest over the central and eastern plains, where it exceeds 1.5 mm² h⁻² in MM (Fig. 4b, c). Warm-season precipitation in this region is dominated by organized convective systems that initiate over the Rockies and propagate eastward overnight (Houze, 2004), so an inability to reproduce the intensity of these systems maps directly onto a deficit in total variance. East Asia separates the two product groups more clearly than CONUS: MS remains broadly consistent with the observations across China, whereas MM still exhibits underestimation over southeastern China. The spatial RMSE summarizes this contrast, at 0.67 mm² h⁻² for MS against 0.95 mm² h⁻² for MM. For the rat_{subD}, both MS and MM underestimate the observations across nearly the entire valid domain. Spatial RMSE values are 10.94% for MS and 22.31% for MM (Fig. 4d, e), with MM showing both a larger magnitude and a more spatially extensive deficit. The ratio of the MM to MS spatial RMSE rises from 1.4 for σ^2_{total} to 2.0 for rat_{subD}, indicating that reanalysis products depart from the observations more in the frequency partitioning of the variance than in its total magnitude."***

➔ Two further paragraphs follow the description of Fig. 6a (601–611).

Lines 609-619: ***"Both MS and MM underestimate σ^2_{1st} from the GP to the southern CONUS, with MM showing the larger deficit over a spatially more extensive area (spatial RMSE of 0.011 and 0.014 mm² h⁻², respectively), indicating that the nocturnal convective systems noted above are underestimated in their diurnal amplitude and not only in their total variance. Over China, MS overestimates σ^2_{1st} over southern China, while MM overestimates it more extensively across central and southeastern China, even though σ^2_{total} is underestimated there (Fig. 4c). The 1st harmonic is therefore not uniformly underestimated across all gridded products. For the rat_1, both MS and MM overestimate it over most of the validation domain, with spatial RMSE values of 17.21% for MS and 26.61% for MM, and with the largest values over interior and northern China. An overestimated rat_1 indicates that σ^2_{reg} is carried almost entirely by the diurnal harmonic, so the second and higher-order components remain weaker than observed and the diurnal cycle is represented as smoother than it actually is. The GP is the exception, where MM underestimates rat_1 over the same area in which σ^2_{1st} is most underestimated (Figs. 6c, e)."***

2. Several figure sequences throughout the manuscript appear redundant. While they present slight variations of the diagnostic metrics, they ultimately convey the same overarching physical conclusion and seem more of supportive rather than adding new insights. This creates repetitive discussion paragraphs and makes it difficult for the reader to pinpoint the primary takeaway. I strongly advise the authors to prioritize key representative figures for the main text and relocate secondary variations to the Supplementary Materials (Specific comments are listed later).

➔ Based on reviewer's suggestion, we have applied a single rule when deciding what belongs in the main text. A figure is retained if it compares the products against the ground-based observations or against one another in a way that carries the argument forward, and a figure is moved to the Supplementary Materials if it repeats an established result for additional cases, additional thresholds, or additional individual products.

The following material has been moved. The station-sensitivity test for the grid cells in Germany and East China is now in the Supplement, and only the representative cell in the central United States is retained in the main text. The per-dataset maps of diurnal phase and amplitude, which previously occupied eight panels, are now shown in the main text as the Multi-Satellite and Multi-Model composites, with the individual products given in the Supplement. The domain statistics computed without the rat_1 threshold, which repeat the result of the thresholded version, are also in the Supplementary Materials. The frequency distributions of the local peak phase, which formerly took ten panels in the main text, have been moved there as well. The vertical moist static energy profiles at the SGP site, which duplicated the location already covered by the following figure, have likewise been relocated. The distribution used to define the moisture-supply threshold and the mean skew-T profiles at the SGP site are given in the Supplement in the same way.

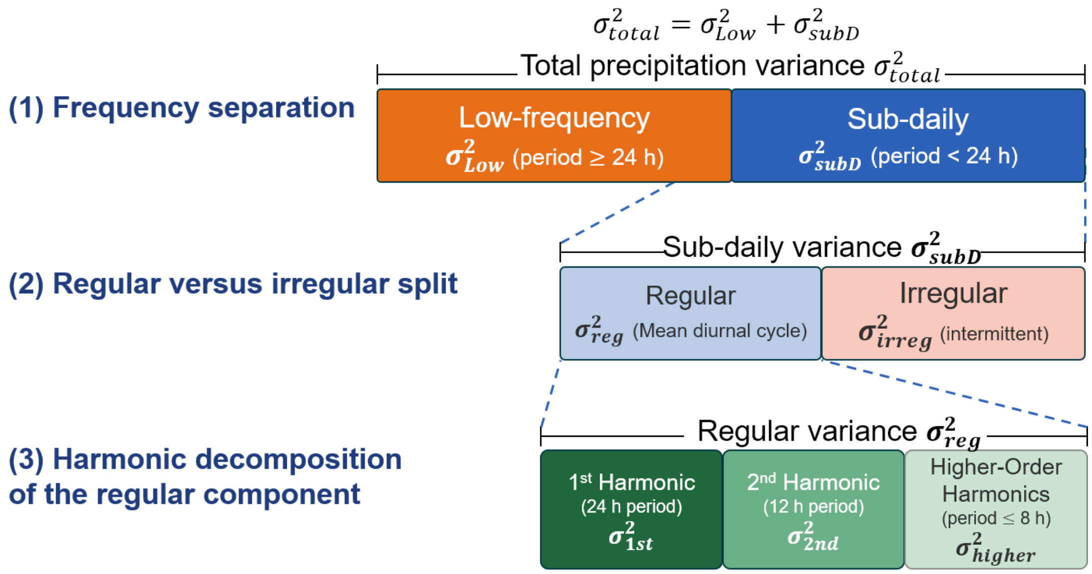
The revised manuscript carries the comparison against the ground-based observations, which the previous version did not include and which is uncommon in global assessments of the diurnal cycle. Figures 4 and 6 now reports the difference between the gridded products and the collocated station observations for the total and sub-daily variance and for their ratio, together with the spatial RMSE of each difference. Figure 9 same is done for the first harmonic variance and its fractional contribution, and for the diurnal phase, where the mean difference and the spatial RMSE are given in each panel. Figure 10 summarises the diurnal phase and amplitude against the ground observations over the quasi-global domain and for the three regions (e.g., CONUS, Europe, and East Asia). Because the argument of the paper is what the products get wrong relative to direct observations, these are the figures we consider representative, and the variations that support them are now in the Supplementary Materials.

Figure 3 and 12 were added rather than moved, and both were requested during the review. Figure 3 is a schematic of variance decomposition, now added to Section 3 so that the hierarchy of the metrics can be followed. Figure 12 replaces the single hemispheric average of the thermodynamic diagnosis with the same diagnosis separated by the moisture-supply regime. Neither repeats an existing figure. This has been revised in the revised manuscript throughout Sections 3 and 4, and the individual relocations are described in our replies to the specific comments that follow.

3. The mathematical decomposition presented in the manuscript involves multiple nested hierarchical tiers - moving from total precipitation variance (σ^2_{total}) to sub-daily variance (σ^2_{subD}), separating regular (σ^2_{reg}) and irregular (σ^2_{irreg}) components, and further decomposing the regular component into Fourier harmonics (σ^2_{1st}, σ^2_{2nd}, σ^2_{higher}). Because of this multi-layered hierarchy and the large number of associated statistical terms and ratios (rat_{subD}, rat_1, $\sigma^2_{1st}/\sigma^2_{total}$), the analytical framework is difficult to follow in pure text form. To improve readability and pedagogical clarity: I strongly encourage the authors to include a methodological flowchart or schematic diagram in Section 2. Providing this visual roadmap early in the paper will make it substantially easier for readers to navigate the subsequent spatial maps and diagnostic discussions in Sections 3 and 4.

- ➔ We appreciate this suggestion. A schematic of the variance decomposition has been added as Fig. 3, showing the three tiers of the hierarchy together with the method applied at each tier and the corresponding equations. It is placed at the end of Section 3.2, where the decomposition and all of its associated ratios have been defined, and therefore precedes every result that relies on these quantities. This has been added in the revised manuscript (Lines 454–455 and Fig. 3).

The full hierarchy of the variance decomposition described in Section 3.1 and Section 3.2 is summarised in Fig. 3."



"Figure 3: Schematic of the sub-daily precipitation variance decomposition, showing the method applied at each tier and the corresponding equations in Sections 3.1 and 3.2. Box widths are uniform and do not indicate the relative magnitude of each component."

4. The final section of the results *'4.3 Physical mechanisms of the diurnal cycle'* is particularly difficult to follow. The authors need to revise this entire section to improve readability and structure. Currently, each paragraph feels disjointed and fluctuates back and forth without arriving at a clear conclusion. Furthermore, the datasets keep changing from one figure to the next throughout the section, which makes it even harder to follow the narrative.

➔ Based on the reviewer's suggestion, Section 4.3 has been rewritten during this revision. The disjointed impression came from the way the section was organised. The previous version moved through the variables one at a time and returned to each product several times, so the reader had to assemble the argument. The section now follows a single line of reasoning. It begins with the global diagnosis, in which the domain is classified according to the magnitude of moisture flux convergence (MFC). The regime classification is described in our response to Comment 32. The analysis then focuses on two-point locations where the proposed mechanism can be evaluated against direct observations. Each part closes with a paragraph that states what has been established rather than leaving it implicit. The global part ends with the finding that the reanalyses reproduce the diurnal phase where precipitation follows the surface energy balance and not where it follows large-scale transport. The SGP section concludes that the three products fail at different stages of the same sequence: NARR reproduces the nocturnal convergence and vertical moisture structure, JRA-3Q and ERA5 capture the convergence but convert it into precipitation about three hours too late, whereas MERRA-2 peaks in the late morning despite moisture being present throughout the column.

Within each part, the discussion now follows a consistent sequence: moisture supply, the thermodynamic state represented by CAPE and CIN, the response of the convective scheme, and finally the precipitation phase. Each product is discussed once following this sequence, rather than being revisited later, so that the schemes are compared at the same stage of the process in each case.

We agree that the changing set of datasets was confusing, and the section now uses a consistent set throughout. CMORPH serves as the observational reference in every figure in Section 4.3, while JRA-3Q, ERA5, and MERRA-2 are included in all analyses. NARR and the IGRA soundings are included only for the two North American sites, because NARR is limited to the North American domain and sounding observations are available only at the SGP site.

The specific changes supporting this restructuring are described in our responses to the corresponding minor comments. These include renaming the section titles, removing the literature summary that previously opened Section 4, separating Fig. 12 according to moisture-supply regime, and moving the secondary figures to the Supplement. These changes are reflected in the revised manuscript in Section 4.3 and Figs. 12–14.

5. Finally, there are numerous acronyms used throughout the manuscript, which makes it challenging for the reader to keep track and forces frequent back-and-forth scrolling. I would strongly suggest defining each acronym at its first occurrence in the main text and, where possible, reducing non-standard abbreviations to improve readability and flow.

➔ The manuscript has been revised in two ways. First, we reviewed the main text, tables, and figure captions to ensure that every abbreviation is defined at its first occurrence and used consistently thereafter. Terms that had previously appeared without definition were either defined or removed.

Second, we applied an explicit criterion to reduce the number of abbreviations. General descriptive terms are abbreviated only when they appear three or more times; otherwise, they are written out in full. This eliminated many abbreviations that had been introduced for only one or two occurrences. Official names of datasets, products, instruments, and data-assimilation systems were treated separately because their abbreviations are used in data tables and figure labels, and repeatedly writing out their full names would reduce rather than improve readability. The abbreviations retained in the main text are therefore either standard terms in the atmospheric and precipitation literature, such as CP, LSP, TP, LST, PMW, LLJ, and MCS, or abbreviations used consistently in figures and tables, such as MS, MM, and the dataset names.

Because the retained set remains relatively large, a complete list of abbreviations and their definitions has been added as Appendix A, allowing readers to identify any abbreviation without searching through the manuscript. These revisions have been implemented throughout the manuscript, with the complete abbreviation list provided in Appendix A.

Abbreviation	**Meaning**
ARM	Atmospheric Radiation Measurement
ARMBEATM	ARM Best Estimate Atmospheric Measurements
CAPE	convective available potential energy

CIN	convective inhibition
CMFD	China Meteorological Forcing Dataset
CMIP	Coupled Model Intercomparison Project
CMORPH	Climate Prediction Center Morphing Technique
CONUS	contiguous United States
CP	convective precipitation
CPC	Climate Prediction Center
DCAPE	dynamic CAPE generation rate
EA	East Asia
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis Version 5
EU	Europe
GP	Great Plains
GPM	Global Precipitation Measurement
GSMaP	Global Satellite Mapping of Precipitation
IGRA2	Integrated Global Radiosonde Archive version 2
IMERG	Integrated Multi-satellite Retrievals for GPM
IR	infrared
JMA	Japan Meteorological Agency
JRA-3Q	Japanese Reanalysis for Three Quarters of a Century
LCL	lifting condensation level
LFC	level of free convection
LLJ	low-level jet
LSP	large-scale precipitation
LST	local solar time
MCS	mesoscale convective system(s)
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, Version 2
MFC	moisture flux convergence
MM	Multi-Model mean
MSE	moist static energy
MSWEP	Multi-Source Weighted-Ensemble Precipitation
NARR	North American Regional Reanalysis
NH	Northern Hemisphere

NOAA	National Oceanic and Atmospheric Administration
PBL	planetary boundary layer
PDF	probability density function
PMW	passive microwave
RMSE	root-mean-square error
SGP	Southern Great Plains
TMPA	TRMM Multi-satellite Precipitation Analysis
TP	total precipitation
TRMM	Tropical Rainfall Measuring Mission
UTC	Coordinated Universal Time
VARANAL	ARM Variational Analysis

"Appendix A: List of abbreviations"

Minor comments

Abstract

1) Please explicitly mention in the Abstract that this study is conducted at a 3-hour temporal scale. Simply referring to a "sub-daily scale" leaves the exact temporal resolution ambiguous to the reader. Abstract:

➔ This is a fair point. The Abstract now states the temporal resolution explicitly, so that the term sub-daily used later in the Abstract is no longer open to interpretation. This has been revised in the revised manuscript (Lines 13–14).

"Analyses are performed at 3-hourly resolution, combining sub-daily variance decomposition and harmonic analysis with atmospheric moisture budget diagnosis across contrasting precipitation regimes."

2) 'Global' to 'quasi-global' as it is not covering the entire global ocean. Also it would better to clarify that this study has been carried out at 60 NS.

➔ We agree that global was imprecise for a domain that excludes latitudes poleward of 60 degrees. Both the term and the latitude range have been corrected in the Abstract. This has been revised in the revised manuscript (Lines 10–13).

"This study evaluates the boreal summer diurnal cycle across five satellite products and four reanalysis datasets against 6,824 ground-based stations over the Northern Hemisphere land regions, complemented by inter-product comparison over quasi-global domain between 60°S and 60°N."

3) I would suggest removing or softening the strong claim in the Abstract stating that PMW upper level ice scattering is the primary driver of the satellite precipitation phase lag. As noted in the main text comments later, algorithm artifacts such as PMW orbital sampling gaps and IR based morphing could be major co-contributors to these timing delays. Rephrasing this in the Abstract to avoid attributing the lag exclusively to a single physical mechanism would be more accurate.

➔ Based on reviewer's suggestion, the Abstract no longer designates ice scattering as the primary mechanism, and now lists the indirect retrieval characteristics together with the algorithmic factors that contribute to the delay (Lines 21–23)

"Satellite products exhibit the observed late-afternoon precipitation maximum with a phase lag of 1–2 hours, arising from the indirect detection of upper-tropospheric ice-scattering signals that persist as convective anvils expand beyond the timing of peak surface precipitation."

Introduction

4) Line 109-130: One of the primary strengths and novel contributions of this study is that it goes beyond simply documenting diurnal variability across estimates to investigate the underlying physical mechanisms driving these patterns. However, this key novelty is currently underemphasized due to how the introduction is structured. While Lines 109–120 effectively lay out the existing literature gaps, the subsequent paragraph (Lines 123–130) jumps directly into procedural steps rather than framing overarching research goals. To better highlight the paper's main novelty and give readers a clear roadmap, I suggest first explicitly stating 2–3 central research objectives that directly address the identified research gaps, followed by a brief outline of the analytical steps taken to achieve them.

➔ We appreciate the reviewer's suggestion. Accordingly, the latter part of the Introduction has been reorganized to present three objectives, each corresponding to one of the gaps identified in the preceding paragraph, followed by a brief description of the datasets and analytical approach (Lines 133–136 and 137–140).

"This study addresses these gaps through three objectives. The first is to establish a ground-based benchmark for how precipitation variability is partitioned across temporal frequencies, quantifying the fraction of total variance that resides at sub-daily scales and the proportion of that variance associated with a diurnal cycle. The second is to evaluate multiple satellite products and reanalyses against this benchmark, extending prior work that has not jointly assessed the diurnal cycle across both product types using a common, dense ground-based reference. The third is to identify the physical origins of the diurnal phase differences among datasets, first by partitioning precipitation into its convective and large-scale components, and then by examining the thermodynamic conditions governing convective initiation in each reanalysis."

"To meet these objectives, we use hourly observations from 6,824 ground-based stations across Northern Hemisphere (NH) land regions. We evaluate sub-daily precipitation characteristics in five satellite products (IMERG; the TRMM Multi-satellite Precipitation Analysis, TMPA; CMORPH; GSMaP; and the Multi-Source Weighted-Ensemble Precipitation, MSWEP) and four reanalysis datasets (ERA5; JRA-3Q; MERRA-2; and the North American Regional Reanalysis, NARR)."

Data

5) What is the main rationale behind selecting those particular datasets. Is there any selection criteria of filters the authors used of only these particular datasets passes or find eligible. If so, please explain briefly that criteria. If no such criteria have been employed briefly mention why these particular datasets have been selected.

➔ We thank the reviewer for this important question. No formal quantitative filtering criteria (e.g., minimum accuracy thresholds) were applied. Instead, the datasets were selected because they represent the products most widely used for diagnosing sub-daily precipitation, and because their design choices span the range of approaches relevant to sub-daily timing, which allows the evaluation to serve as a reference for dataset selection in future work at this time scale.

Among the satellite products, IMERG, CMORPH, and GSMaP all propagate microwave retrievals between overpasses, but differ in how their motion vectors are obtained and in how gauge adjustment is applied: IMERG scales every half-hour within a month by a single ratio, CMORPH matches the daily distribution of its estimates to the gauge analysis, and GSMaP adjusts each hour unevenly so that the 24-hour accumulation matches the daily gauge total. TMPA represents an earlier generation in which calibrated infrared estimates fill the gaps between microwave overpasses. MSWEP differs in kind from the other four, as it merges gauge, satellite, and reanalysis inputs through weighted averaging rather than retrieving precipitation independently.

Among the reanalyses, ERA5, JRA-3Q, and MERRA-2 were selected because they differ in their convective parameterization and trigger formulations, which is the specific quantity of interest for the third objective stated in the Introduction. NARR was included as a contrasting case because, unlike the other three, it directly assimilates precipitation observations. A common requirement across all eight products was a temporal resolution of 3 hours or finer over the analysis domain. A sentence stating the rationale has been added to Section 2.2 (Lines 256–258):

"The satellite products and reanalyses evaluated in this study were selected to span the range of design choices relevant to diurnal precipitation characteristics, so that assessing them against direct ground-based observations provides a reference for selecting among them at this time scale."

6) Considering TRMM has long been deprecated and superseded by GPM, what was the motivation behind including this legacy dataset in the intercomparison? Please add a brief justification in the Data section explaining why TMPA was selected alongside IMERG.

➔ We thank the reviewer for raising this important point. We note that, in response to related comments from this review round, TRMM 3B42RT has been replaced with the research-grade TMPA (3B42) product throughout the revised manuscript, and all corresponding results have been updated accordingly.

➔ TMPA is no longer produced, but it is retained here for what its evaluation reveals about algorithmic evolution rather than for operational relevance. Because IMERG was developed as the direct successor to TMPA and retrospectively extends the GPM multi-satellite framework into the TRMM era, evaluating the two together provides a natural test case for how the representation of the diurnal cycle has changed across algorithm generations. Specifically, TMPA fills temporal gaps between microwave overpasses using calibrated infrared estimates, whereas IMERG, CMORPH, and GSMaP instead propagate microwave retrievals forward and backward in time. This distinction allows the effect of temporal propagation on sub-daily variability to be isolated, and TMPA's markedly different performance in Section 4.1 and 4.2 (e.g., its 1st harmonic variance well above the observed value) indicates that this design choice has a substantial effect on diurnal cycle fidelity. This justification has been added to the revised manuscript (Lines 281–284).

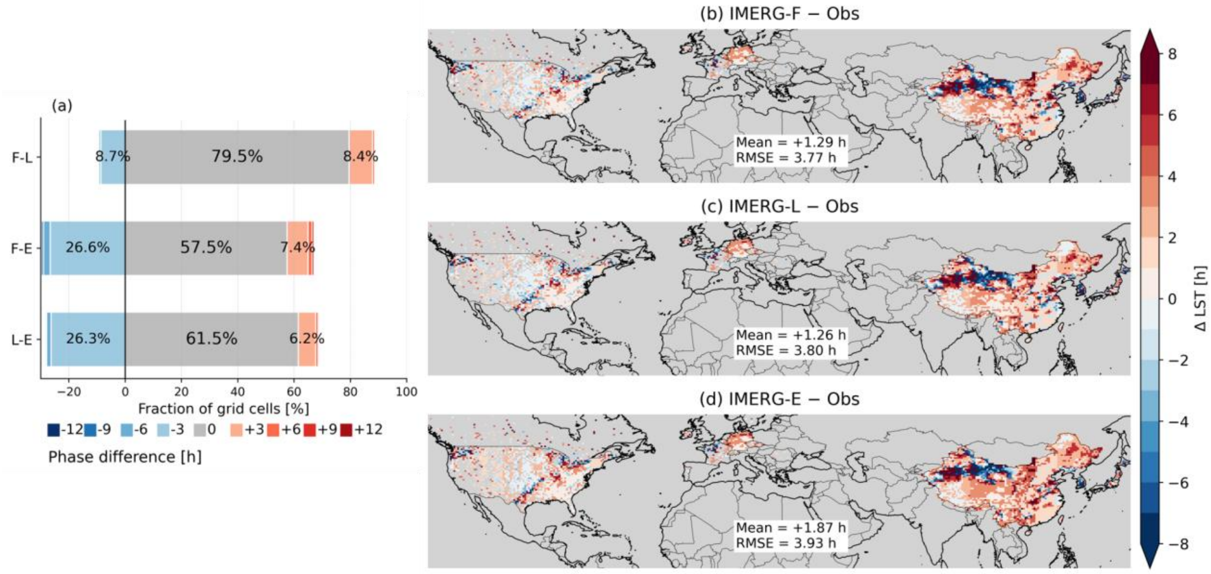
"TMPA is retained despite its discontinued production because, as the predecessor to IMERG, it fills temporal gaps using calibrated infrared estimates rather than propagating microwave retrievals in time. This contrast with IMERG, CMORPH, and GSMaP allows the effect of temporal propagation on sub-daily variability to be assessed across algorithm generations."

7) Moreover, why was the TRMM Realtime version selected here instead of the gauge-corrected one, whereas gauge-corrected versions were selected for the rest of the satellite products? If the idea was to verify the difference between near-real-time and gauge-corrected satellite estimates, wouldn't selecting both products from the same satellite framework make more sense, such as comparing IMERG-F with IMERG-E or IMERG-L.

➔ We thank the reviewer for pointing this out. The revised manuscript in this revision round now uses the gauge-adjusted research product, TMPA 3B42 Version 7, rather than the near-real-time 3B42RT product. The full analysis was reprocessed using 3B42, and all figures, tables, and statistics involving TMPA were recomputed accordingly. This change ensures a more consistent comparison because all products included in the satellite-product analysis now incorporate gauge adjustment. The description in Section 2.2 and the corresponding entry in Table 2 has also been revised accordingly

➔ In addition, following the reviewer's suggestion, we performed a separate sensitivity analysis using the IMERG Final (F), Late (L), and Early (E) Runs over the common 2010–2020 period. All three products were processed using the same regridding procedure and the same harmonic analysis as in the main analysis. Bellow Figure shows the first-harmonic phase differences relative to the ground-based observations. IMERG-F and IMERG-L yield nearly identical domain-mean phase biases of +1.29 and +1.26 h, respectively, with corresponding RMSEs of 3.77 and 3.80 h. IMERG-E shows a somewhat larger delay, with a mean phase bias of +1.87 h and an RMSE of 3.93 h. This description is added to the revised manuscript (Lines 750–754).

"Backward propagation brings the phase closer to the observations, since the IMERG Early (IMERG-E) Run, which propagates forward only, peaks later than the Late (IMERG-L) and Final (IMERG-F) Runs, which propagate in both directions, although a positive offset remains in all three (Fig. S5)."



"Figure S5: (a) Distribution of pairwise phase differences among IMERG-F (F), IMERG-L (L), and IMERG-E (E) over the common observation grid cells. Gray indicates identical phases, while blue and red indicate negative and positive phase differences, respectively. (b–d) First-harmonic phase differences relative to the ground-based observations for IMERG-F, IMERG-L, and IMERG-E."

➔ To further quantify the differences among the three IMERG Runs, we compared their phase differences at the common observation grid cells. For each Run, the phase error was calculated as the circular difference from the observed phase, and the errors were then compared pairwise.

IMERG-F and IMERG-L show particularly close agreement, with identical phases at 79.5% of the evaluated grid cells and a domain-mean pairwise difference of only −0.03 h. In comparison, the differences involving IMERG-E are more pronounced and are also very similar between F–E and L–E. The phase remains identical at 57.5% of the grid cells for F–E and 61.5% for L–E, while the remaining differences show a clear tendency for IMERG-E to produce a later phase than the other two Runs.

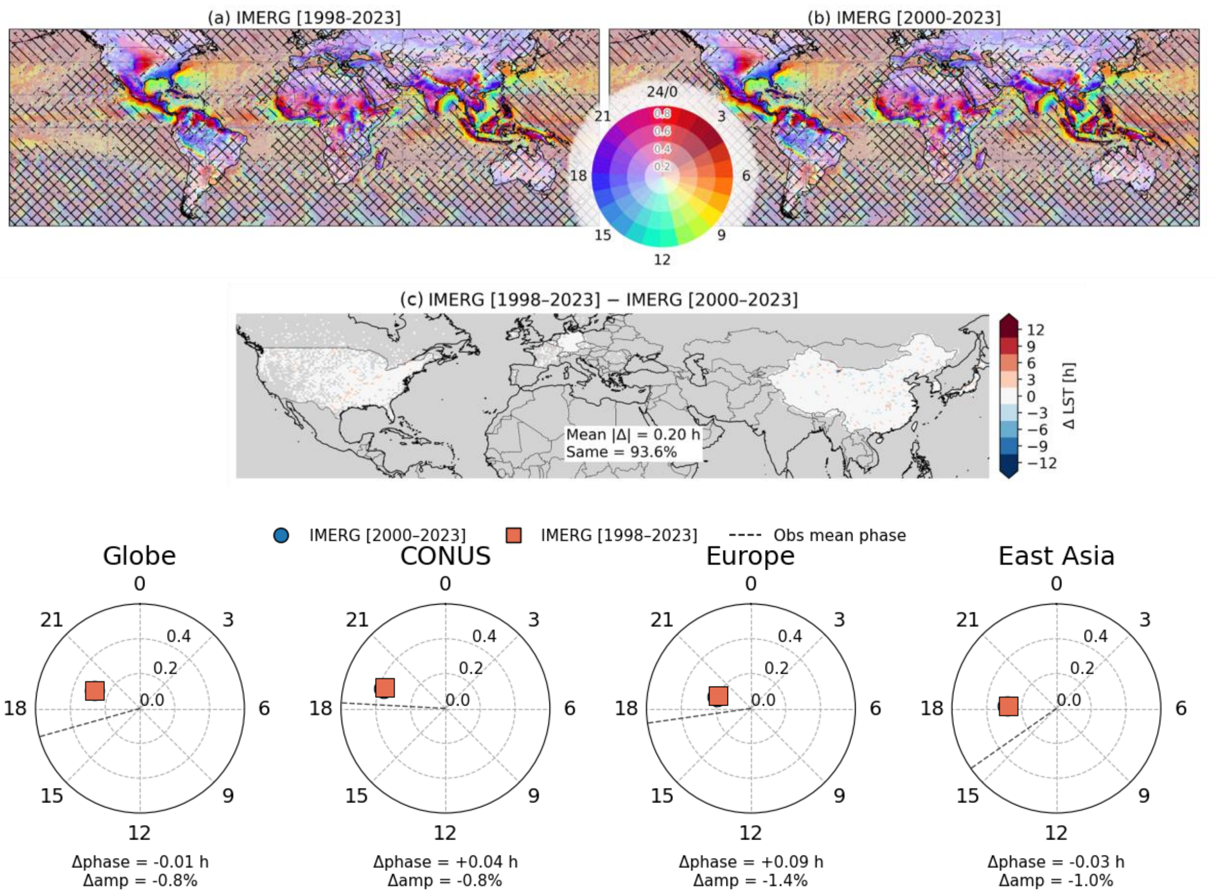
This behavior is consistent with differences in the temporal propagation schemes used in IMERG. The Early Run relies on forward propagation only because of its short latency, whereas the Late Run uses both forward and backward propagation, allowing precipitation estimates to be constrained by microwave observations on both sides of the analysis time. The longer latency of the Late Run can also allow additional microwave observations to become available. The Final Run further incorporates monthly gauge adjustment, for which a spatially varying but temporally fixed monthly ratio is applied to the half-hourly estimates. Because the same factor is applied throughout the month, this adjustment primarily affects precipitation magnitude rather than the relative sub-daily timing.

The close agreement between IMERG-F and IMERG-L, together with the comparable differences of F–E and L–E, therefore, indicates that the larger phase delay in IMERG-E is more consistent with differences in temporal propagation and microwave-data availability than with gauge adjustment itself. More importantly, all three IMERG Runs retain a positive phase offset relative to the observations. This sensitivity analysis therefore indicates that the delayed satellite precipitation phase identified in this study is robust to the choice of IMERG Run and is unlikely to be an artifact of using the gauge-adjusted Final Run.

8) Also, haven't IMERG products been available back to 1998? Even though the authors mention that the record length won't impact the results, this statement seems contradictory to Line 380, where they state: We therefore utilize the maximum available period for each dataset to maximize statistical robustness.

IMERG V7 now extends back to 1998. The version available when the data were retrieved for this study began in 2000, but we have since reprocessed IMERG over its full record, and the analysis now covers 1998-2023. All IMERG results, figures and statistics have been recomputed accordingly, and Section 3 and Table 2 have been updated.

The two periods have nearly identical results. Comparing the 2000–2023 and 1998–2023 IMERG analyses, the mean diurnal phase differs by only 0.01 h and the amplitude by 0.8%. The differences also remain small across the three regions, with phase differences of less than 0.1 h and amplitude differences of less than 1.5%. Thus extending the IMERG record by two additional years has a negligible influence on the derived diurnal characteristics and does not affect any of the conclusions.



9) Also, why did the authors limit the study domain to 60°N–60°S? Is this due to data availability issues, concerns regarding dataset quality beyond 60°N/S, or because diurnal cycle and sub-daily variations are less pronounced beyond 60°N/S, making this region the primary focus of interest?

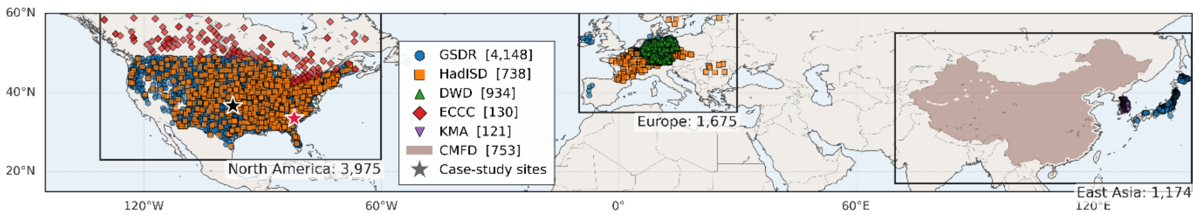
➔ The 60°S–60°N domain restriction is not based on an assumption that the diurnal cycle or sub-daily variability is weaker beyond this range. Rather, it reflects a data availability constraint applied specifically to the precipitation-type analysis in Fig. 11, to ensure a consistent comparison across all datasets shown there. CMORPH's native coverage is limited to 60°S–60°N, consistent with the coverage of several other satellite products. The reanalysis datasets extend further poleward, but retaining latitudes where the satellite products are unavailable would leave the intercomparison undefined in that region.

10) Line 153: I expected Table.1 to be placed immediately underneath the text, as it was cited earlier in the text then of Fig.2.

➔ Table 1 has been moved so that it now follows the paragraph in which it is first cited. In the previous version, it appeared at the end of Section 2.1.1, after both Fig. 1 and Fig. 2, which place it well away from the point of citation. This has been revised in the revised manuscript (Line 176).

11) Although the case study location and metadata are described in the text (Data section), they are missing from the figures and tables. Because the case study represents a major focus of the results, it would be much clearer for the reader if its boundaries were explicitly marked in Figure 1 (e.g., with a bounding box or inset map) and its key specifications included in a summary table rather than remaining text-only.

➔ The two case-study sites are now marked in Fig. 1 with stars, with their coordinates provided in the caption, so that their location relative to the station network is readily identifiable. A summary of the site specifications has also been added as Table 4, introduced at the point where the two sites are first analyzed. The table provides the coordinates, the analysis period, and the precipitation, surface heat flux, and upper-air observations available at each site, along with the precipitation reference and the datasets evaluated there. It also clarifies that ground-based observations are not available for the southeastern U.S. site, where CMORPH is instead used as the precipitation reference and the atmospheric variables are compared among the reanalysis datasets. We placed Table 4 in Section 4.3 than the Data section because it is intended to directly contrast the two case-study locations and their available reference information at the point in the manuscript where they are analyzed, complementing rather than duplicating the site description already provided in Section 2.1.2. This has been revised in the revised manuscript (Fig. 1, Table 4, and Lines 918–920).



“Fig. 1: Locations of ground-based precipitation observations over the Northern Hemisphere. Symbols denote point-based observation stations from each dataset, and the shaded area indicates the spatial coverage of the gridded CMFD dataset, which is a fusion product

combining station observations with reanalysis data. Bracketed numbers indicate the total number of stations or the number of stations to construct the gridded CMFD product. Boxes delineate the three major regions of station concentration (North America, Europe, and East Asia), with the total number of observations within each region indicated below each box label. Stars mark the two case-study sites examined in Section 4.3, the Southern Great Plains site (black; 36.6025°N, 97.4925°W) and the southeastern U.S. site (red; 33.5°N, 82°W).”

	Southern Great Plains	**Southeastern U. S.**
Coordinates	36.6025°N, 97.4925°W (Lamont, Oklahoma)	33.5°N, 82°W (Augusta, Georgia)
Period	JJA, 2000–2023	JJA, 2000–2023
Precipitation observation	ARMBEATM hourly gauge	None
Surface heat fluxes observation	VARANAL	None
Upper-air observation	IGRA2 radiosonde (USM00074646)	None
Precipitation reference	Gauge observation	CMORPH
Evaluated datasets	JRA-3Q, ERA5, MERRA-2, NARR	

“Table 4: Specifications of the two case-study sites examined in Section 4.3. The observations are available only at the Southern Great Plains site, so CMORPH serves as the precipitation reference at the southeastern U.S. site. The locations of both sites are marked in Fig. 1.”

“The two cases are examined next at the point scale using the ARM SGP site in the U.S. GP, where nocturnal precipitation dominates, and a southeastern U.S. site characterized by afternoon precipitation. The sites and the measurements available at each are summarized in Table 4.”

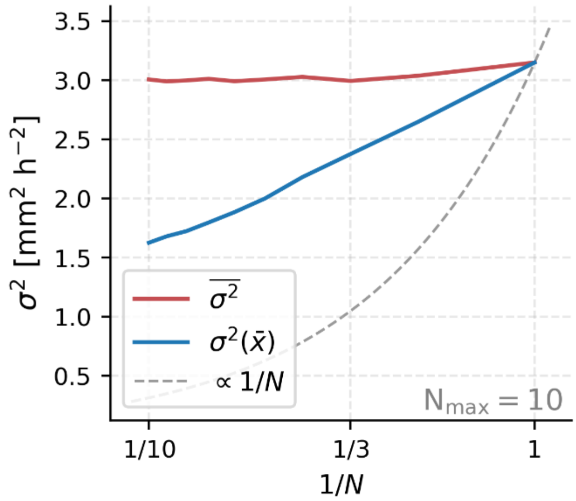
12) Line 196-209: The sensitivity analysis evaluating the 'variance-then-average' method is thorough and appreciated. However, to improve the conciseness and readability of the main manuscript, I recommend moving Lines 196–209 and Figure 2 to the Supplementary Information, leaving just a brief sentence and reference in Section 2.

➔ The passage has been shortened and the supporting material moved to the Supplement. Fig. 2 now shows a single representative grid cell in the central United States, with the corresponding result for Germany and East China given as Fig. S1, and the description in Section 2.1.1 has been reduced to the reasoning behind the choice and a brief statement of what the figure shows. We have retained one panel in the main text because the order of operations justifies underlies every variance statistic reported in Section 4. This has been revised in the revised manuscript (Lines 210–218, Fig. 2, and Fig. S1):

“Sub-daily precipitation variance is computed independently at each ground-based station

prior to aggregation within each grid cell (hereafter, variance-then-averaging approach). An alternative approach, in which station time series are first averaged at each time step across all contributing stations within a grid cell before variance is computed from the resulting composite time series (hereafter, averaging-then-variance approach), would introduce a systematic dependence on the number of contributing stations (N), since the variance of the mean of random variables is proportional to 1/N of the individual variance. Grid-cell variance estimates from that approach would therefore be biased as a function of station density, an attenuation documented in prior studies (Osborn and Hulme, 1997; Light et al., 2022). The variance-then-averaging approach adopted here follows the "statistic-then-grid" order of operations employed in widely used observational datasets (Donat et al., 2013; Gibson et al., 2019).”

“Figure 2 confirms this behaviour at a representative grid cell in the Central United States, where station subsets of increasing size N are randomly drawn from the full set of N_{max} available stations. The variance-then-averaging estimate ($\overline{\sigma^2}$) remains approximately constant across all values of N, whereas the averaging-then-variance estimate ($\sigma^2(\overline{x})$) decreases monotonically as N increases, at a rate more gradual than the theoretical 1/N scaling because neighbouring stations are positively correlated. The same behaviour is found at grid cells in Germany and East China (Fig. S1).”



“Figure 2: Sensitivity of the grid-cell averaged sub-daily precipitation variance (σ^2) to the number of stations (N), evaluated at a representative 0.5° × 0.625° grid cells in the Central United States (33.0°N, 96.9°). For each N from 1 to N_{max}, all possible combinations of station subsets are used to compute σ^2 under two approaches: ($\overline{\sigma^2}$), in which variance is computed at each station individually and subsequently averaged across the selected stations (red line); and $\sigma^2(\overline{x})$, in which station time series are first spatially averaged and variance is subsequently computed from the resulting mean time series (blue line). The dashed line indicates the theoretical 1/N scaling expected when variance decreases solely as a result of averaging independent samples.”

Results

13) The section titles '4.2 Diurnal cycle patterns and underlying mechanisms' and '4.3 Physical mechanisms of the diurnal cycle' are repetitive and confusing. I suggest renaming them to clearly distinguish between their focus areas to make the structure more reader friendly.

➔ Based on the reviewer's comment, Section 4.2 has been renamed to reflect what it presents, which is the evaluation of diurnal phase and amplitude across the datasets, and Section 4.3 retains its title as the section where the physical mechanisms behind those differences are diagnosed. The distinction between the two sections is therefore evaluation followed by mechanism rather than mechanism in both. This has been revised in the revised manuscript (Line 680).

"4.2 Diurnal phase and amplitude across datasets"

14) Use of Sub-subsections: To improve readability and navigation, I suggest introducing subsubsections in Section 4. Currently, the narrative flows across large, uninterrupted paragraphs, making it difficult to discern where the analysis shifts. For instance, creating distinct subsections for the Irregular Sub-daily Variance (Virreg, Fig. 3) and the Regular Sub-daily Variance (Vreg, Fig. 4) would provide a clearer structural hierarchy and help readers follow the results much more easily.

➔ Sub-subsections have been introduced at each point where the focus of the analysis shifts. Section 4.1 is now divided into "4.1.1 Partitioning of the total variance", which covers the sub-daily and lower-frequency components, the partitioning into regular and irregular components, and the sub-daily variance ratio, and "4.1.2 Harmonic composition of the regular component", which examines the first, second, and higher-order harmonics and their contributions. This structure separates the overall variance partitioning, including the irregular component, from the subsequent detailed analysis of the regular component. We did not introduce a separate sub-subsection for the irregular component because it is discussed only briefly within the variance partitioning analysis, whereas the regular component is examined in detail across several figures.

Section 4.3 is likewise divided into "4.3.1 Moisture-supply regimes over Northern Hemisphere land" and "4.3.2 Case studies at two sites", separating the domain-wide diagnosis from the point scale analyses. Section 4.2 is retained as a single section because the analysis forms a continuous evaluation of diurnal phase and amplitude, progressing from spatial patterns to regional statistics and then to the convective and large-scale precipitation components. This has been revised in the revised manuscript (Section 4, Line 502, Line 558, Line 835, and Line 837).

"4.1.1 Partitioning of the total variance"

"4.1.2 Harmonic composition of the regular component"

"4.3.1 Moisture-supply regimes over Northern Hemisphere land"

"4.3.2 Case studies at two sites"

15) Lines 474–478: Please consider removing the opening sentences of Section 4. Restating literature background and research gaps here is redundant with the introduction and creates unnecessary distraction at the start of the Results section. Removing these sentences will allow Section 4 to launch directly into the observational findings and keep the narrative concise.

➔ The opening three sentences of Sections 4.1 have been removed, and the section now begins with the observed total variance. The material they contained is already covered in the introduction. This has been revised in the revised manuscript (Line 517).

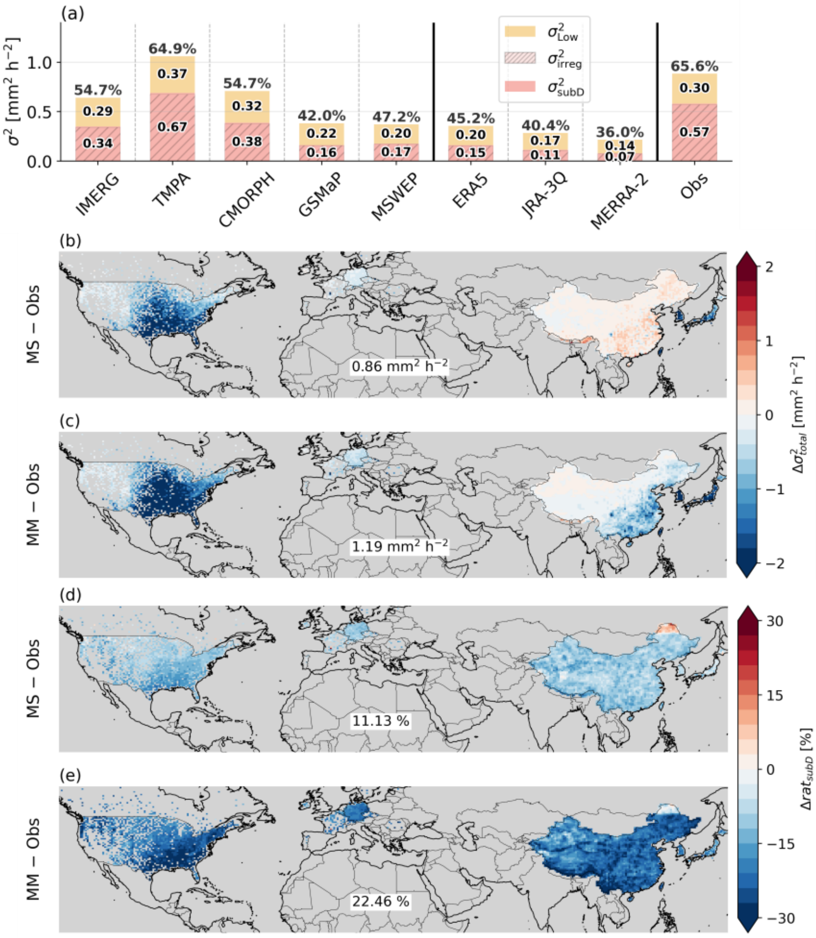
16) Line 478: The statement 'Total variance of precipitation at observational sites during boreal summer is 0.9 $mm^2 h^{-2}$...' is ambiguous. It is unclear whether this value represents an average across all validation sites in the Northern Hemisphere land domain or a specific sub-region. Furthermore, the text does not cite where this specific value appears in the Figures. Please explicitly state the spatial domain being referenced and cite the exact Figure panel/sub-panels (e.g., Fig. 3a/b) at the start of the paragraph. This need for precise Figure citations applies across the entire Results section whenever numerical values are introduced.

➔ The spatial domain and the figure reference have been added at the start of the paragraph. Based on the reviewer's comment, we have also checked the Results section for numerical values introduced without a figure citation and supplied the missing references. This has been revised in the revised manuscript (Lines 511–512).

"Total variance of precipitation at observational sites over NH land during boreal summer is 0.87 $mm^2 h^{-2}$ (Fig. 4a), whereas most of precipitation datasets underestimate this variance, and the underestimation is more apparent in the reanalysis products."

17) Figure 3: Please enhance the legibility of Figures 3a and 3b. The numerical values overlaid inside the bar plots are currently very small and low-contrast, making them difficult to read. Increasing the font size or placing the numerical labels above the bars (or in an adjacent summary table) would significantly improve visual clarity.

➔ The numerical labels in Fig. 3a have been enlarged, and their colour has been changed from white with a dark outline to dark with a white outline, which raises the contrast against the light bar fills. The percentage above each bar has been darkened for the same reason, and the axis tick labels in both panels have been enlarged. We have kept the component values inside the bars rather than moving them above, since the sub-daily variance percentage already occupies the space above each bar and stacking three labels there would crowd the panel. This has been revised in the revised manuscript (Fig. 4a and Fig. 6a).

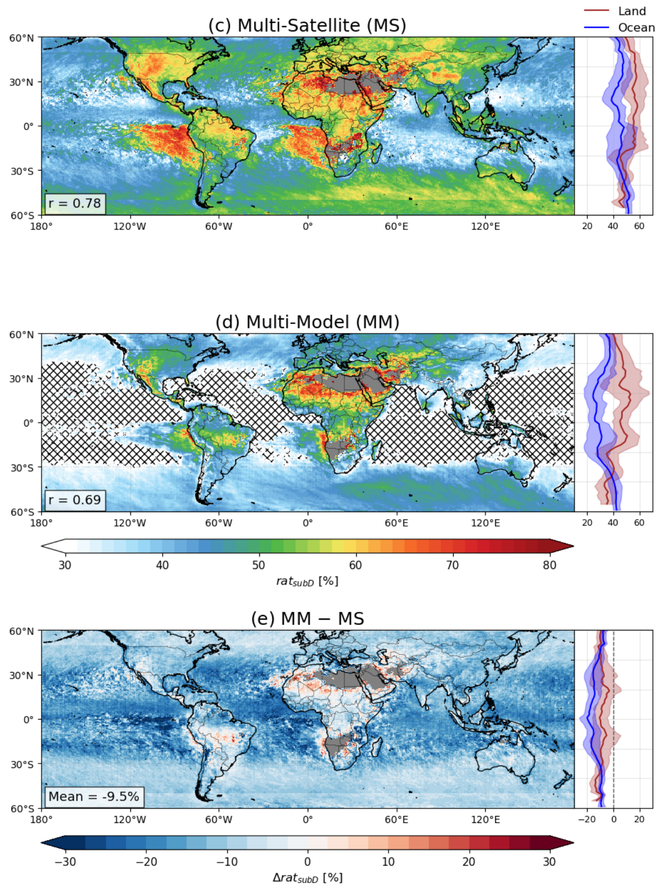


"Figure 4: (a) Sub-daily (pink) and lower-frequency (orange) variance components of precipitation at observation locations. Sub-daily frequency variance is obtained via high-pass Butterworth filter. Numbers inside bars indicate variance magnitude mm^2 h^{-2}, and percentage above bars represent the sub-daily variance ratio (rat_{subD}). Differences from the ground-based observations, for (b, d) the Multi-Satellite (MS; IMERG, TMPA, CMORPH, GSMaP, MSWEP) and (c, e) the Multi-Model (MM; ERA5, JRA-3Q, MERRA-2) mean, in terms of (b, c) total variance (σ^2_{total}) and (d, e) rat_{subD}, where the spatial averaged value is noted in each panel below."

18) In Figure 3c, d, the authors present spatial distributions of ratsubD for the Multi-Satellite (MS) and Multi-Model (MM) ensembles, using spatial correlation coefficients (r) against ground observations as a summary metric. For example, while CMORPH appears closest to ground observations in the global bar plot (0.0091 vs. 0.0079 mm2 h−2), it is impossible to tell their regional distribution. To substantiate the manuscript's core claims regarding ground truth, the authors should include a direct spatial difference maps against gauge observations over land domains with sufficient gauge coverage (e.g., CONUS, Europe, East Asia) rather than relying solely on global mean aggregations. Explicitly mapping the spatial differences would immediately highlight geographical hot spots where reanalysis convective parameterizations succeed or fail (e.g., complex topography vs. flat plains vs. monsoon regions), providing much greater physical utility than global correlation values alone.

In this revision, we include the spatial difference map of MS-Obs and MM-Obs in the manuscript, which were previously displayed only as a spatially averaged bar plot. There was a comment regarding this above, and I intended to include more of the differences from ground-based observation in the main text.

We also compared two product suites directly, shown below. MM lies below MS over almost the entire domain, by 9.5% on average, and the sign does not change between land and ocean or between regions. The difference is therefore systematic rather than confined to particular regimes such as complex topography or monsoon regions, and it is already quantified in the main text by the departures from the observations in Fig. 4 and by the spatial correlations in Fig. 5, so we gave not added a further panel or a separate description.

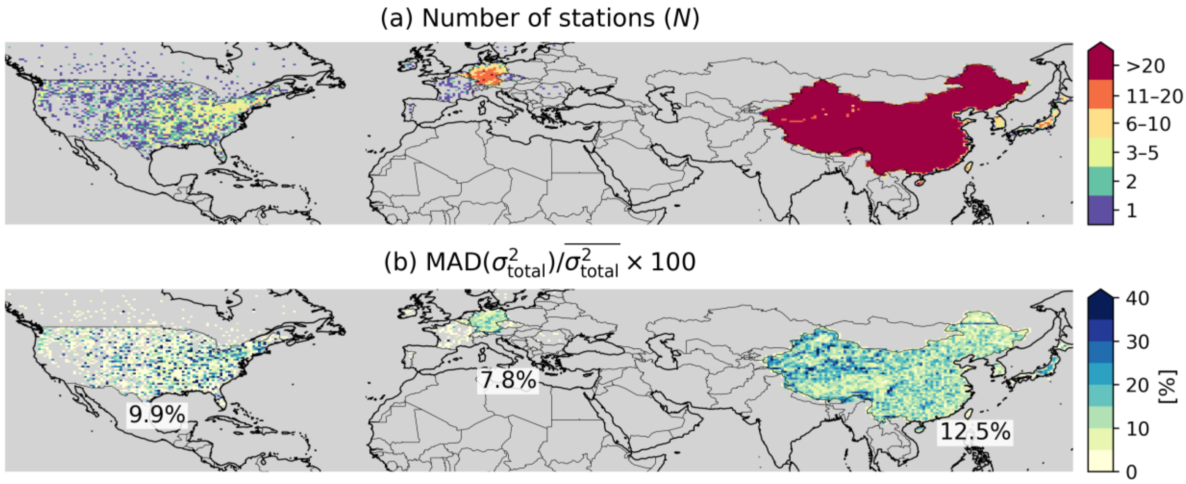


19) Could the authors present a brief regional consistency check (e.g., selecting representative gauge clusters across CONUS, Europe, and China) to examine the variance spread among 3 to 4 nearby stations? Demonstrating this local gauge-to-gauge variability will establish a baseline for observational uncertainty, helping readers better contextualize the performance differences seen in model and satellite estimates.

➔ The check has been carried out, and the results are presented in Fig. S2. The figure shows the number of ground-based stations within each grid cell and the normalized mean absolute deviation of station-level total precipitation variance from the grid-cell mean. Only cells with at least two contributing stations are included in the second panel.

The deviation amounts to 9.9% of the grid-cell mean over North America, 7.8% over Europe and 12.5 over East Asia. This is not related to the number of contributing stations, since Europe provides the lowest values despite having a comparable station density to East Asia. Instead, it appears to reflect the spatial heterogeneity of precipitation within each region. A paragraph reporting these values has been added to Section 2.1.1. This has been included in the revised manuscript (Lines 221–224)

"The spread in precipitation variance among stations within each grid cell indicates how well the grid-cell mean represents station-level values (Fig. S2). The mean absolute deviation remains below 13% of the grid-cell mean and shows no systematic relationship with the number of stations."



"Figure S2: (a) Number of ground-based stations (N) contained in each 0.5°×0.625° grid cell. (b) Mean absolute deviation (MAD) of the station-level total precipitation variance ($\sigma^2_{total,i}$) from the grid-cell mean, normalised by that mean variance. The MAD is calculated as the mean absolute difference between each $\sigma^2_{total,i}$ and the corresponding grid-cell mean variance, where $\overline{\sigma^2_{total}} = \frac{1}{N}\sum_{i=1}^{N}\sigma^2_{total,i}$. Number indicate the regional averages over North Americal, Europe, and East Asia.

20) Line:530: "higher spatial correlation (Δr=0.09)" is this a typo?

➔ The value is not a typographical error. It denotes the difference between the spatial correlations of the Multi-satellite mean and the Multi-model mean against the ground-based observations, which are given in Fig. 5b, c. Since the difference notation was not defined in the text, we note state the two correlation values directly. This has been revised in the revised manuscript (Lines 585–587).

"In terms of spatial agreement with ground-based observations over NH land, the MS demonstrates superior performance over the MM, with a spatial correlation of 0.78 compared with 0.69."

21) 660-664: The authors attribute the observed phase lag in satellite estimates primarily to the physical delay between upper-level ice scattering (detected by PMW sensors) and surface rainfall. How can authors be sure this is the primary mechanism driving the lag? Couldn't PMW sampling gaps and the reliance on morphing would have also be playing a major role here, or at least contributing significantly to the observed delay? It would be helpful if the authors could clarify how the impact of algorithm morphing and sampling density was accounted for.

➔ We agree that microwave sampling density and the temporal interpolation methods used by the satellite products may contribute to the phase lag, and we did not decompose these effects in the present analysis. Isolating their individual contributions would require component-level analyses of the microwave snapshots, the propagated fields, and the infrared-based estimates separately.

Regarding morphing, our results constrain its role in two ways. First, propagation carries a precipitation field forward from an earlier overpass and could therefore report it later than it actually occurred; however, TMPA does not propagate microwave retrievals at all, instead filling gaps in microwave coverage with calibrated infrared estimates, and a comparable phase lag is nonetheless present in TMPA. Second, among the products that do propagate retrievals, a weighting scheme favoring the preceding overpass over the following one would shift the estimate toward earlier conditions; however, CMORPH combines forward- and backward-propagated retrievals with weights inversely proportional to the elapsed time from the overpass rather than its direction (Xie et al., 2017), and a comparable lag is present there as well. Neither propagation itself nor a directional preference within it therefore appears to be required for the lag to occur.

Regarding sampling density, we cannot constrain its contribution within the present analysis. IMERG, CMORPH, and GSMaP draw their microwave retrievals from the same satellite constellation, so any sampling-related effect would act on all three products similarly and would be indistinguishable from the detection delay in our comparison. To our knowledge, no study has yet quantified a systematic shift in retrieved diurnal phase attributable specifically to overpass sampling, and we therefore do not exclude a contribution from sampling density to the observed lag.

The text has been revised to present the lag as consistent with a combination of retrieval characteristics and algorithmic sampling and interpolation effects rather than assigning it to a single mechanism. This has been revised in the revised manuscript (Lines 736–737, 745–746, and Lines 747–754).

"This systematic delay is consistent with the indirect detection characteristics inherent in satellite precipitation retrieval algorithms"

"These detection characteristics cannot be separated from the effects of microwave sampling and temporal interpolation in the present analysis."

"The temporal interpolation alone does not account for the delay: TMPA exhibits comparable phasing to the other satellite products despite filling the intervals between overpasses with calibrated infrared estimates rather than propagating retrievals in time, and a comparable lag also appears in CMORPH, whose forward- and backward-propagated retrievals are weighted by elapsed propagation time without preference for either direction (Xie et al., 2017). Backward propagation brings the phase closer to the observations, since the IMERG Early (IMERG-E) Run, which propagates forward only, peaks later than the Late (IMERG-L) and Final (IMERG-F) Runs, which propagate in both directions, although a positive offset remains in all three (Fig. S5). The contribution of sampling

density cannot be constrained in the same way, since IMERG, CMORPH, and GSMaP draw their retrievals from the same satellite constellation, and a sampling effect would act on all three products alike."

22) Lines 671: What is the primary physical mechanism behind this behaviour? Specifically, how are the satellite products able to capture the timing so well over CONUS, and what accounts for the underestimation of phase? Please add some physical discussion here to help explain these contrasting behaviours.

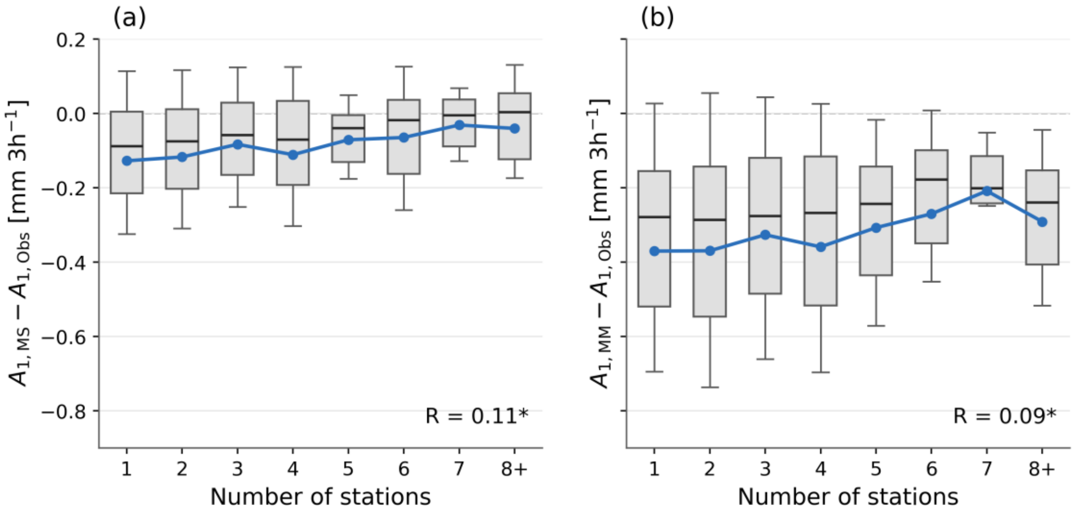
➔ The phase and amplitude describe different properties of the diurnal cycle and are constrained by different aspects of the retrievals. Phase depends on when the maximum occurs, whereas amplitude measures the strength of the diurnal variation, so a product can place the maximum at the observed hour while still representing it as weaker than observed.

Over CONUS, the satellite products reproduce the phase well because the diurnal cycle there has a distinct spatial structure that the retrievals observe directly: precipitation peaks in the afternoon over the western and eastern parts of the domain and at night over the central plains, where mesoscale convective systems propagate eastward from the Rocky Mountains, and the satellite estimates capture this transition. The reanalyses do not reproduce this pattern as well, since their convective schemes respond to surface heating rather than to propagating systems, which is why their phase correlations over CONUS are lower than those of the satellite products.

The amplitude deficit has a different origin, and part of it lies in the ground-based reference itself rather than in the products. Grid cells over CONUS contain fewer stations than those over Europe or East Asia, and a cell represented by a single station carries the diurnal amplitude of that one record, whereas a cell represented by several stations carries an average in which maxima occurring at different hours partly cancel, a character closer to that of a gridded estimate. We examined this dependence over CONUS and added the result as Fig. S7, which shows that the amplitude difference broadly narrows as the number of contributing stations increases in both the multi-satellite and multi-model means ($R = 0.11$ and $R = 0.09$, respectively; both significant at $p < 0.05$, indicated by the asterisk). The relationship is nonetheless weak, and a substantial deficit remains even where several stations contribute, indicating that station coverage accounts for only part of the amplitude difference. This has been added in the revised manuscript (Lines 761–767 and 785–787):

"One possible contribution to this deficit lies in the ground-based reference rather than in the products. A grid cell represented by a single station carries the diurnal amplitude of that record, whereas a cell represented by several carries an average in which maxima occurring at different hours partly cancel, which is closer in character to a gridded estimate, and grid cells over CONUS contain fewer stations than those over the EU or the EA. The amplitude difference broadly narrows as the number of contributing stations increases (Fig. S7)."

"This reflects the distinct spatial structure of the diurnal cycle over CONUS, with an afternoon maximum over the western and eastern parts of the domain and a nocturnal maximum over the central plains, which the satellite retrievals observe directly whereas the reanalyses do not reproduce from surface heating alone."



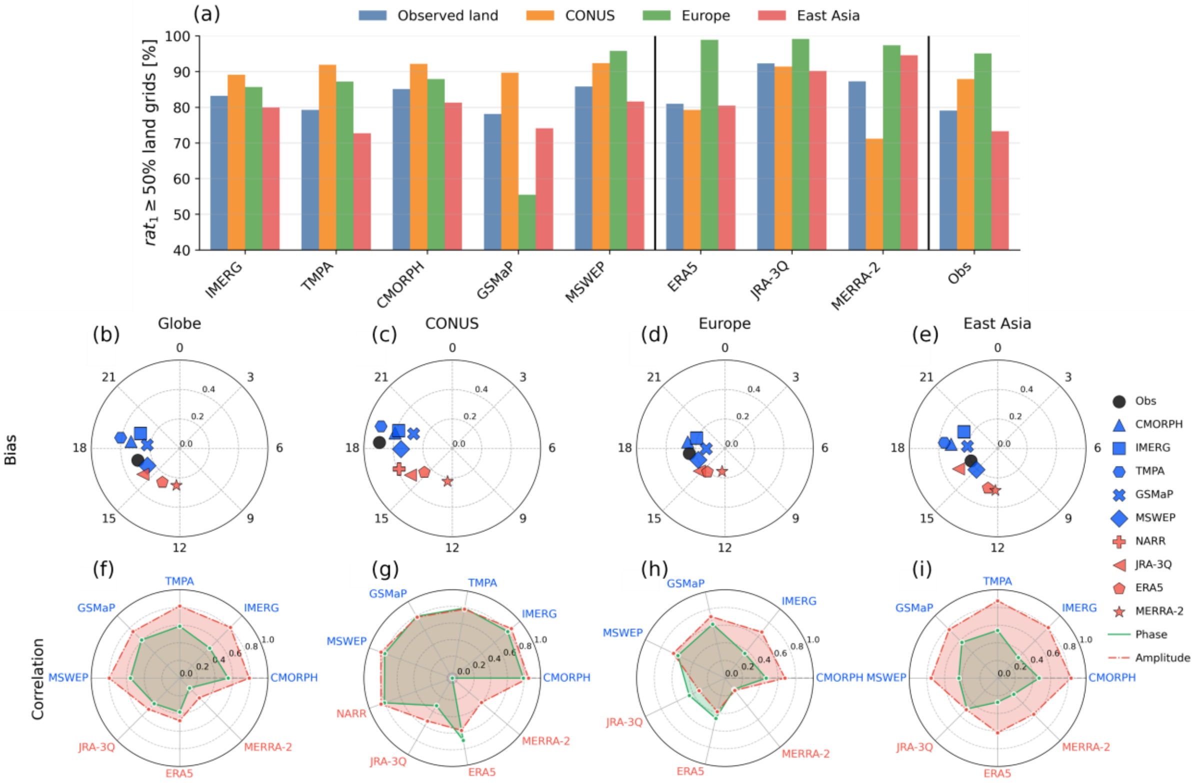
"Figure S7: Differences in 1st harmonic diurnal amplitude between the gridded products and ground-based observations over CONUS as a function of the number of stations contributing to each grid cell, for (a) MS and (b) MM, where grid cells containing eight or more stations are grouped into the 8+ class. Whiskers extend to 0.5×IQR, and the blue line connects the mean difference for each station-count class. R denotes the Pearson correlation between station count and amplitude difference, with an asterisk indicating significance at the 5% level."

Figure. 7

23) Could the author clarify what percentage of global/regional land grid cells meet the $\boldsymbol{rat_1} \geq$ 50% threshold for each dataset.

➔ The threshold is applied to the observations, so a single mask defines the grid cells entering the phase statistics for every dataset rather than each product being masked by its own $\boldsymbol{rat_1}$. The fraction of land grid cells with observations in which the observed $\boldsymbol{rat_1}$ reaches 50% is 79% over the full analysis domain, and 88%, 95%, and 73% over CONUS, the EU, and the EA respectively.

For completeness, we also report below the fraction of land grid cells satisfying $rat_1 \geq 50\%$ computed from each dataset's own $\boldsymbol{rat_1}$ field (i.e., without applying the observation-based mask), as shown in Fig. S6a. Over the full domain, these fractions range from 78% (TMPA) to 92% (MSWEP) across the five satellite products, compared with 79% in the observations; the three reanalyses (excluding NARR, which is evaluated only over CONUS) range from 82% (MERRA-2) to 87% (ERA5). Regionally, most values lie between 70% and 99%, the principal exception being GSMaP over Europe, at 55%. The reanalyses generally exceed the observed fraction over Europe and East Asia, with the exception of MERRA-2, which falls below the observed fraction in all three regions (71% over Europe and CONUS, and correspondingly lower over East Asia).



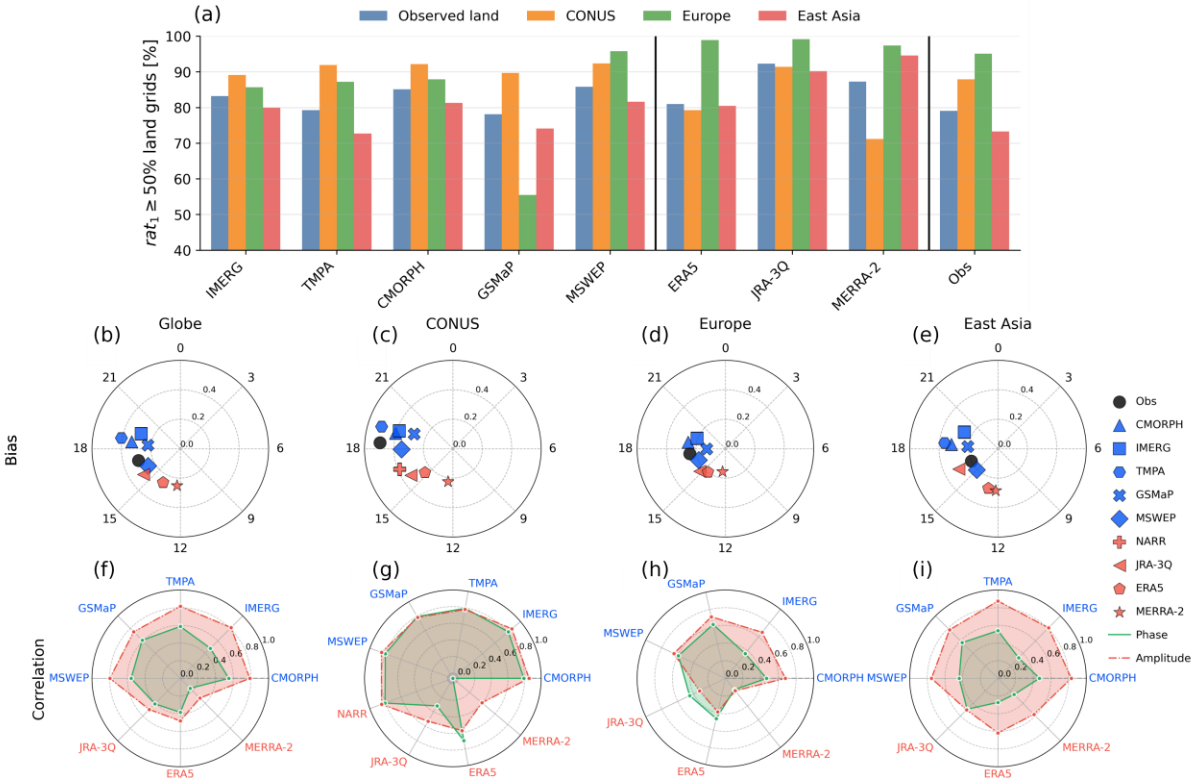
"Figure S6: (a) Percentage of land grid cells with ground-based observations for which $rat_1 \geq 50\%$, given for the observed land domain as a whole (blue) and separately for CONUS (orange), Europe, and East Asia (red). (b-i) Same as Fig. 11, but without the $rat_1 \geq 50\%$ criterion, so that all grid cells with ground-based observations are included."

24) Please consider showing unconditioned phase/amplitude domain statistics (or lowering the threshold) in the Supplementary Information. This would help confirm whether the reported peak phase leads (12–14 LST in reanalyses) are robust across all continental grid cells or if they are sensitive to the chosen threshold

➔ Based on the reviewer's suggestion, we repeated the calculation without the rat threshold and added the result as Fig. S5f–i. The lead of the reanalyses relative to the observations is retained, and their peak phases remain in the early afternoon, so the result reported in the main text is not an artefact of the threshold. The absolute phases shift by less than an hour, the satellite estimates becoming slightly later and the reanalyses slightly earlier, and the amplitudes are smaller in every dataset. These shifts follow from the character of the grid cells that the threshold removes. Because the mask is evaluated on the observations, removing it restores cells in which the observed 24-hour harmonic accounts for less than half of the regular sub-daily variance, that is, cells where the diurnal cycle is weak or poorly defined. The number restored differs by region. Over the EU the threshold was already met at 95% of the observed land cells and the domain statistics barely change, whereas over CONUS and East Asia, where 88% and 73% were retained, the shift is more apparent. The reduction in amplitude reflects the weaker diurnal signal in the restored cells. The threshold is retained in the main text so that the comparison is restricted to cells where the 1st harmonic carries the greater part of the observed regular sub-daily variance, that is where the observed phases and amplitude provide a meaningful reference. The caption of Fig. 10 now refers to Fig. S5 for the unconditioned version.

This has been revised in the revised manuscript (Lines 729–733 and Fig. S3).

"Figure 10: Spatially averaged diurnal phase (azimuth) and amplitude (radius) of the precipitation from ground-based observation (black), satellite (blue), and reanalysis (red) datasets over (a) the globe, (b) Contiguous United States (CONUS), (c) Europe (EU), and (d) East Asia (EA), for grid points where the observed $rat_1 \geq 50\%$. The spatial correlation of diurnal phase (green solid) and amplitude (red dashed) from gridded products with the observation is depicted in (e)-(h). The corresponding statistics without this criterion are shown in Fig. S6."



"Figure S6: (a) Percentage of land grid cells with ground-based observations for which $rat_1 \geq 50\%$, given for the observed land domain as a whole (blue) and separately for CONUS (orange), Europe, and East Asia (red). (b-i) Same as Fig. 10, but without the $rat_1 \geq 50\%$ criterion, so that all grid cells with ground-based observations are included."

Figure. 8

25) It appears CMORPH is being used as the representative product for the satellite ensemble in the subsequent intercomparisons. Please add a brief explanatory line prior to this transition explicitly stating that CMORPH serves as the satellite benchmark, along with a short justification for choosing it over other evaluated products.

➔ The CMORPH is used as the satellite reference from Fig. 11 onwards, and this was not stated before the comparison began. We have added a sentence at that point giving the choice and the reason for it (Lines 803–805):

"Among the satellite products evaluated above, CMORPH shows the closest agreement with ground-based observations in representing global diurnal characteristics. We therefore adopt it as the reference for the grid-scale comparison."

26) Line 704: Are the authors referring to Figure 7 or 8?

➔ Thank you for catching this. The sentence has been rewritten during the restructuring of Section 4.2 and now refers to Fig. 11 (Line 802):

"To identify the physical processes underlying these phase shifts, we examine the respective contributions of the CP and LSP components to the diurnal cycle of TP (Fig. 11)."

27) In the caption of Figure 8, the authors refer to 'backward and forward masking in Fig. 6'. This terminology is somewhat ambiguous. Rephrase text references to use standard descriptive terms such as 'low precipitation/amplitude mask (\ hatching)' and 'harmonic dominance mask (/ hatching)' to avoid ambiguity.

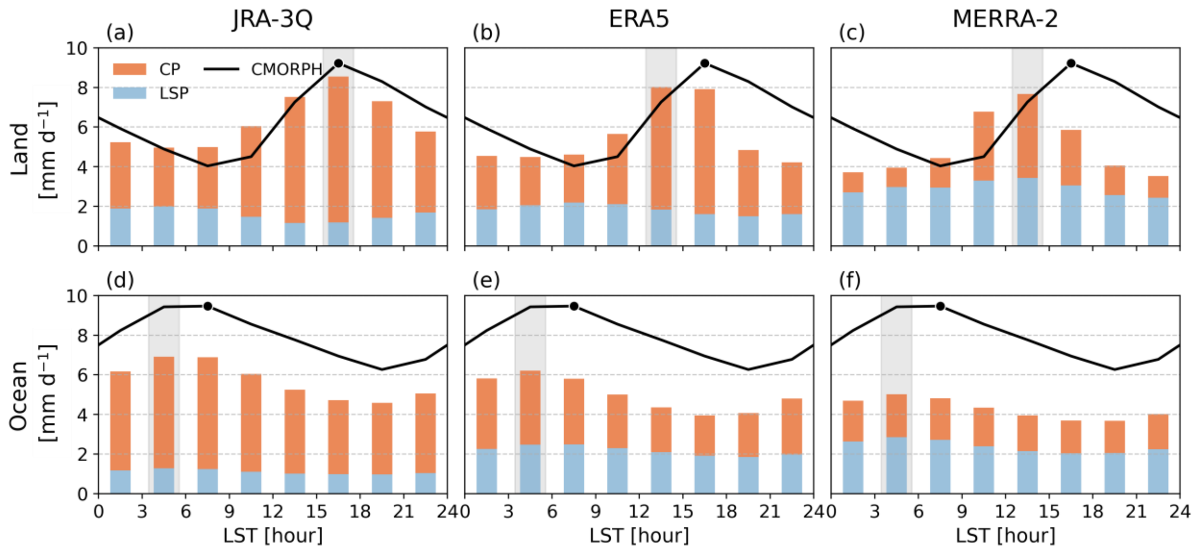
➔ The two masks are now named by their criteria rather than by hatching direction. The naming is given in the caption of Fig. 9, which is the revised form of the former Fig. 6, and the caption that refers to that masking now point that figure. This has been revised in the revised manuscript (Fig. 9 and Fig. 11 captions):

"Figure 9: Global maps of diurnal phase (colour) and amplitude (colour intensity) of precipitation from (a) MS and (b) MM. Backward (\) slashed area denotes the low precipitation amplitude mask, applied where the daily precipitation amount is less than 0.275 mm or normalized amplitude of diurnal cycle < 30% (Watters et al., 2021), and forward (/) slash denotes the harmonic dominance mask, applied where $\mathtt{rat}_1$ is less than 50%. (c, d) Differences in diurnal phase from the ground-based observations for (c) MS and (d) MM, with the mean difference and the spatial RMSE given in each panel. The corresponding maps for the individual datasets are shown in Fig. S3."

"Figure 11: Diurnal cycle of area-averaged precipitation over land (a-c) and ocean (d-f) grid cells bounded by 60°S–60°N (only grid point after the masking depicted in Fig. 9), for JRA-3Q (a, d), ERA5 (b, e), and MERRA-2 (c, f). Bars show the partitioning of total precipitation (TP) into convective precipitation (CP, orange), and large-scale precipitation (LSP, blue) components, and the black line shows the TP from CMORPH. Background shaded bars indicate the local time at which each reanalysis reaches its TP maximum."

28) It would be insightful to see a brief comparison (at least as a supplements) of how well the reanalysis estimates agree among themselves when separating convective from non-convective precipitation. Does the relative contribution of convective vs. stratiform rain show strong interproduct consistency, or do different parameterization schemes yield significant discrepancies?

➔ The comparison suggested here required a quantity that the original Fig. 8 did not provide, since that figure counted grid cells rather than precipitation amounts. We have therefore replaced it with the diurnal cycle of area-averaged precipitation partitioned into its convective and large-scale components, shown against the CMORPH total for land and ocean separately, and moved the frequency distribution to the Supplement as Fig. S8. The partitioning proves to differ substantially among the products at its absolute level. CP accounts for 86% of TP in JRA-3Q over both land and ocean, for 67% over land and 56% over ocean in ERA5, and for less than half in MERRA-2. Through the day, however, the three products behave consistently, in that the diurnal variation of TP over land follows that of CP in every product, so the large-scale component sets the level of the total rather than its phase. Section 4.2 now reports both. This has been revised in the revised manuscript (Fig. 11 and Lines 802–821):

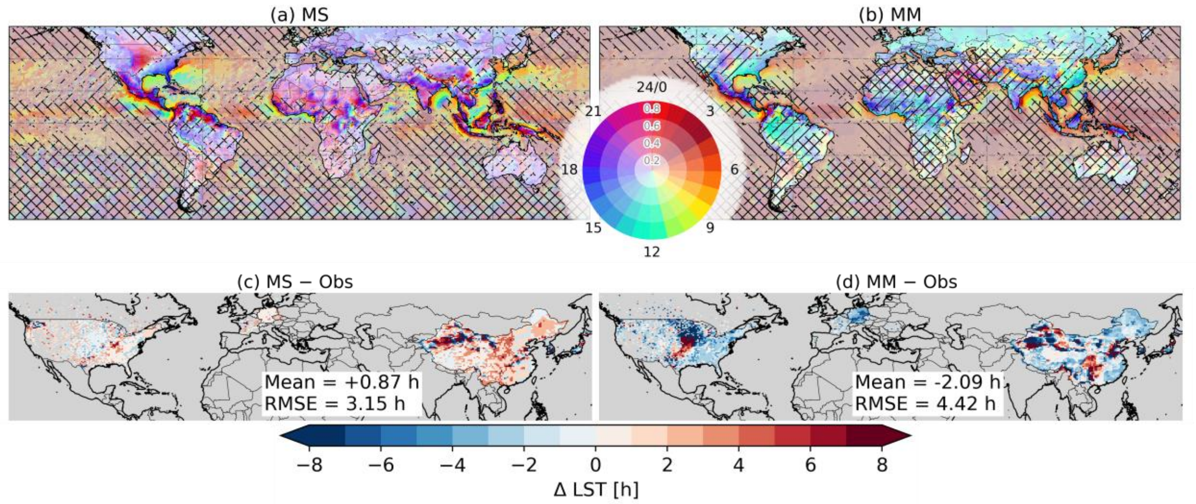


"Figure 11: Diurnal cycle of area-averaged precipitation over land (a-c) and ocean (d-f) grid cells bounded by 60°S–60°N (only grid point after the masking depicted in Fig. 9), for JRA-3Q (a, d), ERA5 (b, e), and MERRA-2 (c, f). Bars show the partitioning of total precipitation (TP) into convective precipitation (CP, orange), and large-scale precipitation (LSP, blue) components, and the black line shows the TP from CMORPH, with the black dot marking its diurnal maximum. Background shaded bars indicate the local time at which each reanalysis reaches its TP maximum."

Figure. 6, 7, and 8: Figures 6, 7, and 8 all focus primarily on the diurnal phase and amplitude (with Figure 8 adding the precipitation type perspective). Because these figures largely repeat the same core message with slight variations, the authors might consider moving some of them (or their redundant panels) to the Supplementary Material to keep the main text concise and focused.

➔ We thank the reviewer for this suggestion, and we have restructured the figures for Section 4 accordingly. The redundancy identified here arose from presenting the same 'phase and amplitude' information once per dataset. In the revised manuscript, these dataset panels are consolidated into two composites, the Multi-Satellite (MS) and Multi-Model (MM) means, and this consolidation is applied throughout Section 4, with the corresponding maps for the individual datasets given in the Supplement (Fig. S3). The former Fig. 6, which comprises eight per dataset maps, is now Fig. 9 with two composite maps, and the former Fig. 8, which comprised ten histogram panels, has been replaced by Fig. 11 with six panels. The number of panels in these two figures is thereby reduced from eighteen to ten.

The space released by this consolidation is used for a comparison that the previous version did not contain. Figure 9 now also shows the difference in diurnal phase between the gridded products and the ground-based observations, with the mean difference and the spatial RMSE reported in each panel (Figs. 9c, d). The three figures consequently no longer convey the same quantity. Figure 9 gives the spatial structure of the phase error against observations. Figure 10 summarises that evaluation quantitatively, using circular statistics (circular mean phase and circular correlation) to place every product and region in a single view. Figure 11 partitions the area-averaged precipitation amount into its convective and large-scale components, which is a different variable and responds to the reviewer's preceding comment. Each therefore answers a different question, namely where the error occurs, how large it is for each product, and what produces it. Because this evaluation is made against the station network rather than among the products themselves, which is not commonly done in global intercomparisons of the diurnal cycle, we would prefer to retain these three figures in the main text while the redundant panels remain in the Supplement. This has been revised in the revised manuscript (Fig. 9, Fig. 10, Fig. 11, Lines 681–687, Lines 729–733)



"Figure 9: Global maps of diurnal phase (colour) and amplitude (colour intensity) of precipitation from (a) MS and (b) MM. Backward (\) slashed area denotes the low precipitation amplitude mask, applied where the daily precipitation amount is less than 0.275 mm or normalized amplitude of diurnal cycle < 30% (Watters et al., 2021), and forward (/) slash denotes the harmonic dominance mask, applied where rat_1 is less than 50%. (c, d) Differences in diurnal phase from the ground-based observations for (c) MS and (d) MM, with the mean difference and the spatial RMSE given in each panel. The corresponding maps for the individual datasets are shown in Fig. S3."

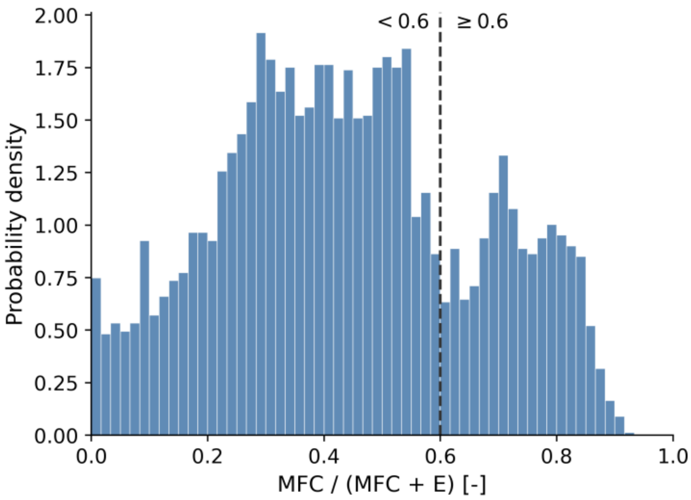
Physical mechanisms of the diurnal cycle:

29) (Figure.9): Domain-averaging thermodynamic variables (CAPE, CIN, MFC, E) across the entire 0∘–60∘N region can conflate contrasting convective mechanisms (such as oceanic convection, afternoon monsoon storms, and nocturnal LLJ systems). It might be more informative for the authors to replace or complement the hemispheric mean in Figure.9 with regional diagnostic breakdowns across a few representative climate zones. This would help highlight how specific parameterizations perform under different regional regimes.

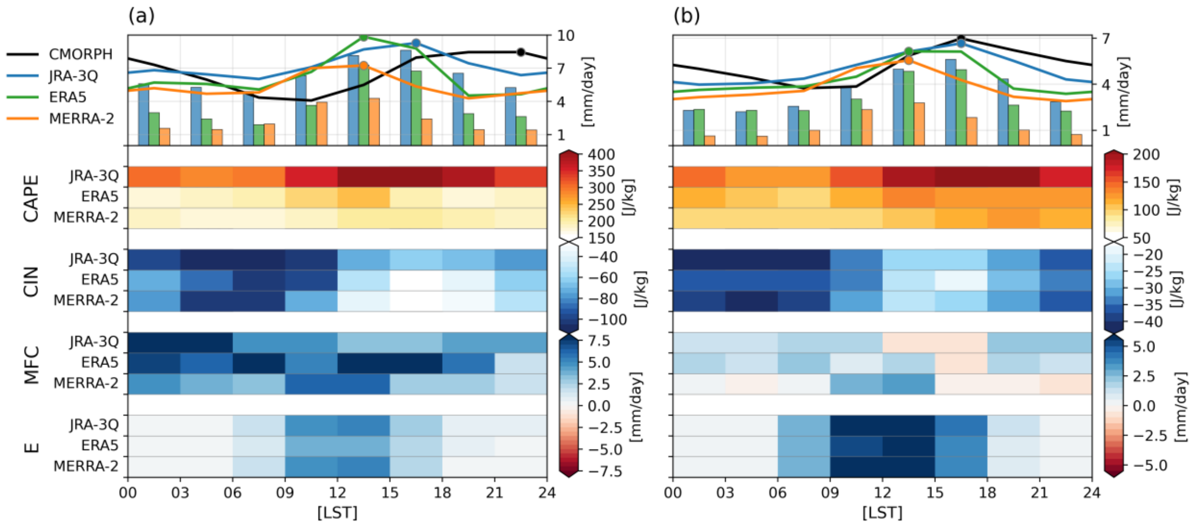
➔ We agree that the hemispheric means mixes regimes that behave differently, and we have changed the figure accordingly. Rather than splitting the domain by climate zone, we split it by the source of the moisture, because that is what the convective schemes respond to. Regions within the same climate zone can be supplied in different ways, so a division based on the moisture supply separates the mechanisms more directly than a geographical one.

For each grid cell we computed the climatological daily mean ratio MFC/(MFC+E), averaged over the three reanalyses. The distribution of this ratio has two modes, and we used the minimum between them, 0.6, was therefore adopted as a threshold (Fig. S9). Grid cells with ratios at or above 0.6 are classified as the convergence-supplied regime, whereas those below 0.6 are classified as the evaporation-supplied regime. Revised Fig. 12 now presents precipitation, CAPE, CIN, MFC, and E separately for these two regimes rather than as a single hemispheric mean.

This separation reveals a distinction that was obscured in the original figure. In the evaporation-supplied regime, CMORPH peaks at 15–18 LST. JRA-3Q reproduces this phase, whereas ERA5 and MERRA-2 peak approximately three hours earlier. In the convergence-supplied regime the CMORPH maximum shifts to 21–24 LST, but all three reanalyses continue to peak in the afternoon, their peaks therefore occur approximately 6–9 hours earlier than in CMORPH. The reanalyses perform differently when convection is driven mainly by daytime surface heating, but all three fail to reproduce the nocturnal maximum associated with sustained moisture convergence and dynamical lifting. This analysis has been added to Fig. 12 and Section 4.3, with the threshold distribution in Fig. S9 (Lines 842–846).



"Figure S8: Probability density of the climatological daily-mean ratio MFC/(MFC+E), averaged across JRA-3Q, ERA5, and MERRA-2, over land grid cells on rainy days. The dashed line marks the threshold of 0.6 used to separate the two moisture-supply regimes."



"Figure 12: Diurnal cycle of area-averaged precipitation and its relevant variables for rainy days (daily precipitation ≥ 0.275 mm day⁻¹) over 0°–60°N, separated by the moisture-supply regime into grid cells where the climatological daily-mean ratio MFC/(MFC+E) is (a) at or above 0.6 and (b) below 0.6. Upper panel shows the diurnal cycle of TP (lines) and CP (solid bars) from CMORPH, JRA-3Q, ERA5, and MERRA-2, with circles marking the peak hour of each dataset. Lower panels show diurnal variations of CAPE, CIN, MFC, and E, where each row represents three reanalysis datasets (JRA-3Q, ERA5, and MERRA-2)."

30) Line 767–768: The authors state that 'CIN remains strongly negative during nighttime hours and begins to erode from 12 LST onward, reaching its minimum around 15 LST.' Looking at the corresponding panel (middle colorbar), there do not appear to be any positive values shown at all. Could the authors please clarify the sign convention being used here for CIN (e.g., whether they are referring to absolute magnitudes or signed values) and ensure the text and colorbar scale are aligned?

➔ CIN is shown as a signed quantity and is negative throughout, which is why no positive values appear in the panel. The text described its weakening during the day, which refers to the magnitude rather than the signed value and makes it ambiguous. We have rewritten the CIN descriptions so that they refer to magnitude consistently. This has been revised in the revised manuscript (Lines 877–878, 945–947, and 998–999):

"CAPE reaches its peaks at 12 LST, three hours earlier than in JRA-3Q, while CIN follows a pattern similar to JRA-3Q, weakening from 12 LST and reaching its smallest magnitude at 15 LST."

"JRA-3Q shows a double-peak structure, with an afternoon peak at 15–18 LST driven by the CAPE maximum as CIN erodes to its smallest magnitude at 12–18 LST and a nocturnal peak (03–06 LST) driven by MFC."

"Convective initiation occurs only after surface heating erodes the CIN to a near-zero between 12 and 18 LST, facilitating the rapid release of stored energy."

31) Another issue with the 0∘N–60∘N average figures and results is that the values appear quite low (e.g., CAPE reaching around 350 J/kg, which makes it hard to believe they represent real convective environments. Are the authors implying that we should interpret these values strictly in relative terms rather than absolute physical magnitudes?

➔ The values are intended to be read in relative terms. They are domain and time averages over land between 0°N and 60°N on rainy days, and this domain includes arid, high-latitude, and elevated regions where CAPE remains small even on rainy days. In addition, the timing of maximum instability differs among grid cells, which further reduces the peak in the spatially averaged diurnal cycle. The use of pressure levels common to all reanalysis products also smooths the vertical thermodynamic structure and limits the interpretation of the absolute CAPE and CIN magnitudes.

The same pressure levels, parcel definition, and conditional sampling are applied to every product. CAPE and CIN are computed with Metpy (May et al., 2022) on the levels common to all four reanalyses, namely 1000, 925, 850, 700, 500, and 400 hPa, with the parcel lifted from the lowest level available at each grid point. What the analysis rests on is the phase relationship between CAPE, CIN and precipitation within each product and the differences in that relationship between products, neither of which depends on the absolute magnitude. We have added the details of the calculation to Section 3.4 and noted in Section 4.3 that the values are used for comparison among the products rather than as absolute measures of instability (Lines 482–484 and 844–845):

"CAPE and CIN are computed with Metpy (May et al., 2022) using a surface-based parcel on the pressure levels common to all four reanalyses, namely 1000, 925, 850, 700, 500, and 400hPa, with the parcel lifted from the lowest level available at each grid point."

"The variables are averaged over the grid cells of each regime, so they are used for comparison among products rather than as absolute measures of instability."

32) Furthermore, a broad 0∘N–60∘N spatial average is highly generic and overly simplistic. Such large-scale domain averages should serve primarily as general reference background rather than the basis for making strong claims about reanalysis performance or convective physics. Please address this dilution effect or qualify these broad claims.

➔ We agree that the original 0°N–60°N hemispheric average was too generic to support specific conclusions about reanalysis performance or convective physics, since it mixes grid cells with fundamentally different moisture-supply mechanisms. To address this, we replaced the single domain average with separate analyses of the convergence-supplied and evaporation-supplied regimes (now Fig. 12), classified using the climatological ratio MFC/(MFC+E) at a threshold of 0.6 (Fig. S9). This separation reveals a distinction that was previously obscured: in the evaporation-supplied regime, the reanalyses differ from one another and from the observations by only up to three hours, and JRA-3Q reproduces the observed phase closely, indicating that the comparison is informative about differences in convective trigger formulations. In the convergence-supplied regime, by contrast, all three reanalyses are indistinguishable from one another and depart from the observations by six to nine hours, because none of their convective schemes represent precipitation initiated by sustained large-scale lifting rather than surface heating. The original hemispheric average obscured this contrast entirely. This has been revised in the revised manuscript (Fig. 12, Fig. S9 and 844–846):

"The variables are averaged over the grid cells of each regime, so they are used for

comparison among products rather than as absolute measures of instability. The separation follows the climatological daily-mean ratio MFC/(MFC+E), averaged across the three reanalyses, at a threshold of 0.6 (Fig. S9).”

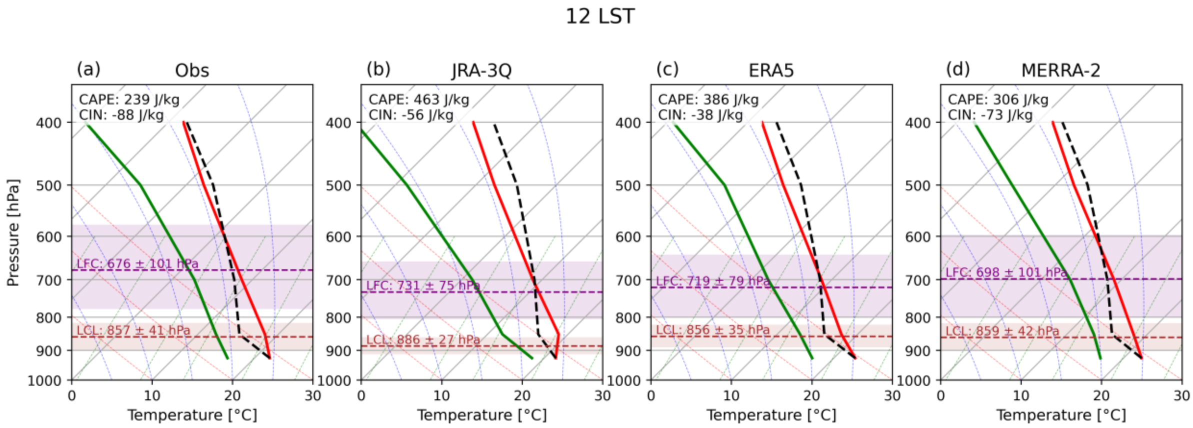
33) Additionally, what explains the large differences in CAPE across the reanalysis products? Is this driven by differences in convective parameterizations, vertical resolution, or parcel definitions (e.g., Surface-Based vs. Mixed-Layer)? Please add a brief explanation in the text.

➔ The CAPE shown here is not the archived CAPE field of each reanalysis. It is computed offline from the pressure-level temperature and dewpoint with MetPy, in the levels common to all products, with a parcel initialized at the lowest available level, and under the same conditional sampling. Differences in the diagnostic itself, including the parcel definitation and the levels entering the integration, therefore cannot explain the spread. Differences in convective parameterization or in native model resolution may still contribute indirectly, by shaping the thermodynamic profiles from which CAPE is computed.

To examine these thermodynamic differences without the dilution associated with the broad-domain average, we additionally evaluated CAPE, CIN, LCL, and LFC at the ARM SGP site (Fig. S10). At 12 LST the CAPE is largest in JRA-3Q (462.9 J kg^{-1}), followed by ERA5 (385.6 J kg^{-1}) and MERRA-2 (306.1 J kg^{-1}). The LFC follows the same ordering, occurring at 731.1, 719.4, and 697.9 hPa, respectively. JRA-3Q also has a higher-pressure LCL (885.9 hPa) than ERA5 (855.7 hPa) and MERRA-2 (859.1 hPa), indicating a distinct lower-tropospheric thermodynamic state.

The LCL and LFC are physically related to the CAPE differences because they affect the parcel thermodynamic path and the depth over which positive buoyancy develops. However, neither level alone determines the CAPE magnitude, which ultimately depends on the vertically integrated parcel buoyancy between the LFC and equilibrium level. Their differences therefore provide additional evidence that the reanalyses produce different thermodynamic profiles. Such differences can arise from the combined effects of data assimilation and model physics, including the boundary-layer and convective parameterizations and native model resolution, and the present analysis does not separate their individual contributions, A brief statement to this effect has been added to the text, and the diurnal cycles of CAPE, CIN, LCL, and LFC are provided as Fig. S10 (Lines 865–867):

“Consistent with this, the LFC ordering at 12 LST follows the CAPE ordering, with JRA-3Q showing a lower LFC and larger CAPE, indicating that differences in thermodynamic structure contribute to the inter-reanalysis spread in CAPE (Fig. S10).”



"Figure S10: Mean skew-T profiles at 12 LST on rainy days at the ARM SGP site (36.6025°N, 97.4925°W), for (a) the IGRA radiosonde observation, (b) JRA-3Q, and (c) ERA5, and (d) MERRA-2. Red and green solid lines show temperature and dew point, and the black dashed line shows the surface-based parcel path. The brown and purple dashed lines mark the mean lifting condensation level (LCL) and level of free convection (LFC), with shading spanning ±1 standard deviation. CAPE and CIN averaged over the same profiles are given in each panel."

34) Also, why do the CAPE values appear so low here? Could the authors clarify how CAPE was computed and provide a brief note on the typical CAPE range expected for this particular site during convective conditions?

➔ As clarified in our response to the preceding comment, the relatively low CAPE values result from both the calculation method and the spatial/temporal averaging applied here. CAPE is calculated using a surface-based parcel on the pressure levels common to all products, namely 1000, 925, 850, 700, 500, and 400 hPa, with the ascent integrated only to 400 hPa. The resulting values are therefore smaller than those calculated from a full-resolution atmospheric profile. They are also averaged over all rainy days during boreal summer rather than composited over active convective events, and over broad moisture-supply regimes rather than individual grid cells, which further reduces the apparent magnitude as described in our response above.

Under this sampling and calculation framework, the broad average substantially reduces the apparent CAPE magnitude because it combines regions with very different convective environments and different timings of maximum instability. The distribution across individual grid cells shows that CAPE commonly reaches several hundred J kg^{-1}, with the median-to-upper portion of the distributions spanning approximately 200–800 J kg^{-1} across the reanalyses. The domain-mean values shown in Fig. 12 are therefore used only as a climatological reference for comparing the diurnal evolution of CAPE, CIN, and precipitation among the products, rather than as representative values of individual convective environments or as a basis for claims about absolute convective intensity. We have clarified this limitation in Sections 3.4 and 4.3.

35) Line 783: Could the authors clarify what is meant here by 'large-scale dynamical forcing'? Please be specific by providing concrete physical examples (e.g., synoptic-scale vertical motion, frontal lifting, or thermodynamic barriers like CIN) so the reader understands the exact mechanism being referenced.

➔ "Large-scale dynamical forcing" originally referred to the same processes introduced earlier in the manuscript, moisture convergence and free-tropospheric lifting, but the use of a different phrase for the same processes obscured this connection. We have replaced the term with an explicit description of what the DCAPE-based trigger conditions on, and we have expanded the discussion of the corresponding mechanism in ERA5 for clarity.

Specifically, the DCAPE trigger in JRA-3Q conditions convective onset on the rate of CAPE generation by these named processes (moisture convergence and free-tropospheric lifting), in addition to a CAPE threshold, which holds convection back through the transitional period described in the Introduction (Lines 57-61), during which shallow cumulus growth is limited by entrainment of dry environmental air before deep convection can be sustained.

“This delay is attributed to the initial development of shallow cumulus, which experience lateral and cloud-top entrainment of dry environmental air into the convective plumes. This process dilutes the buoyancy within the core of the rising air and suppresses its vertical growth (Stirling and Stratton, 2012; Tian and Kuang, 2016). Consequently, a transitional period is required for shallow convection to moisten the lower free atmosphere and further erode CIN.”

In ERA5, by contrast, the modified Tiedtke scheme's closure relaxes only the free-tropospheric part of CAPE, subtracting the contribution generated by boundary layer forcing so that the full instability becomes available to deep convection only at the end of the lower-tropospheric heating and moistening cycle (Bechtold et al., 2014). The phase alignment we find between the CAPE maximum and precipitation onset suggests that this adjustment does not fully hold convective onset back through this transition, and this residual early bias is consistent with prior evaluations of the closure itself, which reported an early bias of up to two hours over the eastern United States. This has been revised in the revised manuscript (Lines 873–876 and 877–885):

“By conditioning the onset on both thermodynamic instability and its rate of generation, the trigger holds convection back through the transition in which shallow cumulus, limited by entrainment of dry environmental air, moisten the lower free troposphere and further erode the residual CIN.”

“In ERA5 the two conditions are separated. CAPE reaches its peak at 12 LST, three hours earlier than in JRA-3Q, while CIN follows a pattern similar to JRA-3Q, weakening from 12 LST and reaching its smallest magnitude at 15 LST. TP rises with the CAPE maximum and forms a plateau that holds between 12 and 18 LST, so precipitation is generated before CIN is fully eroded. CP nonetheless peak at 15–18 LST, the same as CMORPH, so the early rise is carried by the total rather than by the convective component. The closure of the modified Tiedtke scheme (Bechtold et al., 2008, 2014) relaxes only the free-tropospheric part of the CAPE, subtracting the contribution generated by boundary layer forcing so that the full instability becomes available to deep convection only at the end of the lower-tropospheric heating and moistening cycle. Where the destabilization originates at the surface, as in this regime, that subtraction holds the convective onset back far enough for CP to peak with the observations, although it does not prevent precipitation from beginning while CIN is still eroding.”

36) Lines 800–825: It is again very hard to follow what the authors are trying to convey here. By the end of this paragraph, there is no clear synthesis or takeaway; it reads more like a simple list of sequential events rather than a meaningful physical interpretation. Please either rewrite this section to clearly articulate the main conclusion or remove it if it does not add substantial value to the manuscript.

➔ We agree that the original discussion largely described a sequence of thermodynamic changes without a sufficiently clear synthesis. We have therefore removed the original Fig. 10 and its accompanying discussion from the main text. The vertical moist static energy (MSE) profiles remain useful supporting evidence for the SGP analysis and have been moved to the Supplement as Fig. S11. Only the findings directly relevant to interpreting the precipitation phase have been retained and incorporated into the discussion of the revised Fig. 13, which now concludes by identifying the specific processes responsible for the differences among the reanalysis products and their broader implication for reanalysis development. This has been revised in the manuscript (Lines 962–977):

"The three products fail at different points in the same sequence: NARR reproduces both the nocturnal MFC and the vertical moisture structure delivered by the low-level jet, and its precipitation peaks with the observations; JRA-3Q and ERA5 reproduce the nocturnal convergence but convert it into precipitation three hours late, without the mid-tropospheric moisture that the low-level jet should deliver; and MERRA-2 peaks in the late morning, with moisture present at every level but little of it converted into precipitation at all. Precipitation timing at the SGP site therefore depends on three factors: the representation of nocturnal MFC, the fidelity of mid-tropospheric moisture transport by the low-level jet, and the efficiency with which accumulated instability is converted into CP (Song et al., 2013; Xie and Zhang, 2000). That NARR alone reproduces all three is consistent with the constraining effect of precipitation assimilation via latent heating nudging, which converts observed precipitation into vertical latent heating profiles and incorporates them into the temperature tendency equation, thereby constraining the convective and grid-scale schemes to produce precipitation at observed times (Lin et al., 1999; Mesinger et al., 2006). This result indicates that where convective parameterizations fail to represent the physical pathway connecting moisture transport to precipitation, direct assimilation of precipitation observations can compensate for that structural limitation. Beyond the SGP site, NARR exhibits the smallest phase biases and the highest correlations with observations across the broader CONUS domain (cf. Fig. 10). For representative sites in the southeastern U.S. where IGRA observations are unavailable, NARR is therefore used as the reference for diagnosing the diurnal precipitation characteristics of the other reanalysis products."

37) Line:813: The text states, 'CAPE reaches its diurnal minimum (approximately 200 J kg−1) at 06 LST', but looking at Figure 10, the minimum appears to occur at 18 LST instead. Could the authors please check this line and correct the timing if this is a typo?

➔ CIN We re-examined the underlying data and confirmed that the IGRA CAPE minimum occurs during 06–09 LST. The CAPE values at 18 LST are higher, so the timing reported in the original manuscript was not a typographical error. Nevertheless, following the restructuring described in our response to the preceding comment, the original Fig. 10 and the sentence in question have been removed from the main text. The MSE profiles are now provided as supporting information in Fig. S11.

38) Lines 815–819: Does this imply that Convective Inhibition (CIN) plays no significant role here? Since the CIN temporal patterns and values look virtually identical throughout the diurnal cycle, please clarify its relevance to the triggering mechanism or address why CIN remains flat while the diurnal cycle evolves.

➔ This does not imply that CIN plays no role at this site, but rather that its contribution is a persistent, nearly constant background constraint rather than the primary driver of the diurnal timing. CIN and CAPE are plotted on the same axis in Fig. S11, and CIN's amplitude is small relative to that of CAPE, which is why its diurnal variation appears comparatively flat. We agree that the original wording overstated CIN's role by attributing the afternoon suppression of convection specifically to the erosion of residual CIN.

At this site, the diurnal phase is primarily governed by the evolution of CAPE rather than by the removal of inhibition. CAPE reaches its minimum between 06 and 09 LST and increases through the day to a maximum at 00 LST, coinciding with the nocturnal precipitation peak, whereas CIN remains close to constant, exerting a persistent but non-varying constraint throughout the diurnal cycle. We have revised the passage accordingly. This has been revised in the revised manuscript (Lines 938–940)

"Its magnitude varies little through the day, however, so at this site the diurnal phase is governed more by the evolution of CAPE than by the removal of inhibition. Observed CAPE reaches its maximum at 00 LST, which is aligned with the nocturnal precipitation peak."

39) Line 835 (Figure 11): It is quite confusing why another figure is introduced for the exact same location (the ARM SGP site) right after Figure 10. If Figure 11 focuses on the same site, why aren't these panels combined into a single figure, or why isn't one moved to the Supplementary Materials? Please clarify the distinct physical insight provided by Figure 11 or consolidate it with Figure 10.

➔ We thank the reviewer for this observation, and we agree that presenting two consecutive figures for the same site was redundant. We have therefore removed the original Fig. 10 from the main text and moved the vertical MSE profiles to the Supplement as Fig. S11.

The original Fig. 11 has been retained in the main text (now revised Fig. 13) because it provides a distinct type of information: whereas the former Fig. S11 showed the vertical structure of moist static energy as a diagnostic of moisture transport by the low-level jet, Fig. 13 presents the diurnal evolution of precipitation together with CAPE, CIN, moisture flux convergence, and evapotranspiration, which together allow the convective triggering mechanism itself to be diagnosed rather than only the moisture supply that precedes it. The revised main text now uses the MSE profiles (Fig. S11) only as supporting evidence when interpreting the moisture-transport-related results shown in Fig. 13, removing the repetitive discussion of the nocturnal precipitation phase that previously appeared in both figures.

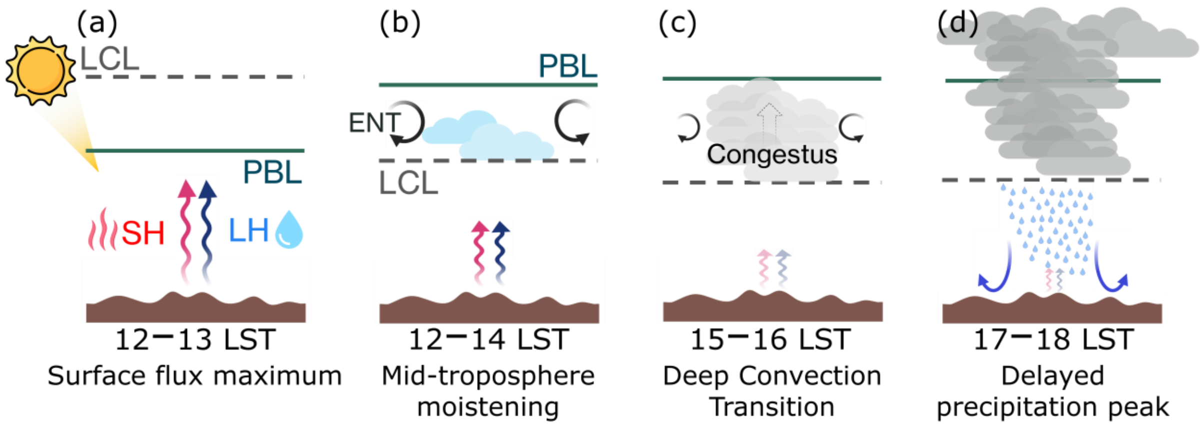
Conclusion:

40) Streamlining the Conclusion Section (Lines 918–961): The section title should be simplified from 'Summary and Conclusion' to simply 'Conclusion'. Furthermore, lines 918 to 961 currently read more like an extended discussion section rather than a concise synthesis. I recommend condensing this text to focus strictly on the key physical findings and actionable take-home messages. Removing detailed narrative re-evaluations from this section will make the core conclusions far more distinct and impactful for the reader.

➔ We agree, and the section has been retitled "Conclusions" and substantially streamlined. The previous version followed the Results section too closely and therefore read as an extended summary. The revised section is instead organized around the main findings and their physical implications. It first establishes the observational benchmark, showing that sub-daily variance accounts for 65.6% of total precipitation variance and is dominated by the irregular component, while all evaluated products underestimate this contribution. It then synthesizes the two product classes, relating the delayed satellite peaks to retrieval and sampling characteristics and the premature reanalysis peaks to the representation of convection. The product-by-product re-evaluations have been removed, retaining for each product only the single attribute that explains its departure from the observations.

The physical interpretation is now carried by a new schematic (Fig. 15) illustrating the shallow-to-deep convective transition over land: from the surface flux maximum and boundary layer growth, through shallow cumulus moistening of the lower free troposphere under dry-air entrainment, to congestus development and deep convection. The observed three-to-six-hour lag between the surface flux and precipitation maxima is interpreted as reflecting this intermediate transition, whereas the reanalyses initiate deep convection before this entrainment-limited moistening stage is completed.

The revised section closes with two main implications. JRA-3Q shows the smallest phase shift in the evaporation-supplied regime, consistent with its trigger accounting for the rate of CAPE generation as well as its magnitude, while NARR reproduces the nocturnal maximum at the SGP site through precipitation assimilation. Together, these results point to two complementary pathways for improvement: more physically realistic convective triggering, and greater use of sub-daily precipitation assimilation. This has been revised in the manuscript (Section 5, Fig. 15, and Lines 1058–1074 and 1081–1089)



"Figure 15: Schematic of the shallow-to-deep transition of convective precipitation over land on a summer afternoon. (a) Surface sensible (SH) and latent (LH) heat fluxes peak near 12–13 LST as the PBL deepens toward the LCL. (b) Shallow cumulus form once the PBL top exceeds the LCL and moisten the lower free troposphere, limited by the entrainment (ENT) of dry environmental air at cloud top. (c) Congestus clouds deepen further within the moistened layer near 15–16 LST, with weaker ENT as the clouds broaden. (d) Deep convection develops thereafter, and precipitation, reinforced by downdraft outflow at the surface, reaches its maximum near 17–18 LST."

"JRA-3Q departs least from the observations, consistent with its DCAPE-based trigger, which conditions convective onset on both the CAPE threshold and the rate of CAPE generation. That constraint holds the onset through the intermediate stage, but only where moisture is supplied locally at the surface. Comparing the evaporation- and convergence-supplied precipitation regimes shows that this advantage is regime-dependent: JRA-3Q reproduces the observed phase where precipitation is evaporation-supplied, whereas in the convergence-supplied regime all three reanalyses retain their afternoon maxima while the observations peak at night, and the trigger formulation no longer separates the products. The ARM SGP site is one such convergence-supplied location, where NARR uniquely reproduces the observed nocturnal precipitation maximum through direct assimilation of precipitation observations via latent heating nudging. Improving reanalysis diurnal cycles therefore appears to depend on two things: convective triggers that respond to large-scale forcing and not only to local instability, and wider use of precipitation assimilation where the trigger alone is not enough.

The sub-daily fluctuations examined here make up most of total precipitation variance, and every product evaluated in this study underestimates them, whether the precipitation is retrieved from satellite or generated by model physics. Daily and longer averages, which most precipitation evaluations still rely on, do not resolve this part of the variability, even though it is what matters for land-atmosphere coupling (Seo and Dirmeyer, 2022) *and for the convective processes analyzed above. Benchmarking several independently developed satellite and reanalysis products against a common set of ground observations helps identify where each product falls short at this time scale and why.*

This study is limited to Northern Hemisphere land regions during boreal summer, where long-term sub-daily gauge records are dense enough to support the evaluation. Extending the analysis to the Southern Hemisphere, tropical regions, and additional seasons, where the forcing mechanisms and dominant precipitation regimes differ from those examined here, remains a priority for establishing whether the biases identified here generalize beyond the regime sampled. The evaluation of convection-permitting reanalyses, which explicitly resolve mesoscale convective dynamics rather than relying on parameterization, offers a particularly promising avenue for determining whether higher-fidelity process representation can reduce the phase and intermittency biases identified here. Additionally, the rapid emergence of AI-based forecast models such as ECMWF's Artificial Intelligence Forecasting System (AIFS; Lang et al., 2024) raises the question of whether data-driven approaches inherit the phase biases of their training data or instead learn to reproduce the observed diurnal cycle more faithfully."