We thank the reviewer (Dr. Randall Donohue) for these helpful comments. Reviewer comments are listed below, along with our response to each. In some cases, we describe the proposed revisions to the manuscript (with line numbers) or refer to proposed revisions described in our responses to the other reviewers, but we recognize that the revised manuscript is requested in a subsequent step.

Comment 1:

I have eagerly read this manuscript and am welcoming of a new perspective on the Budyko curve. While the theoretical understanding of the curve has been continuing to grow, my perspective is that actual improvements to the performance of the model in hydrological applications stalled a few decades ago (my own work included!). It would be great to kick-start that process again. So I am glad to read a critical appraisal of the use of the Budyko curves.

Response 1:

We thank the reviewer for his interest in our manuscript and appreciate the kind words.

Comment 2:

I have one comment to make on this manuscript, which relates to the claim that catchments in reality don't follow Budyko-like curves. The authors show this using catchment data from the US and the UK (Figure 2). The data show that the catchments do generally follow a curve when using the long-term averages but not when using time-series (looking at a catchment through time). This is an expected result due to the underlying assumption in the framework that precipitation is the *only* supply of water. This is often interpreted in terms of catchments needing to be in steady state.

Given this inherent model assumption, one should expect time-series catchment data NOT to follow a single curve. Hence, when the authors say that

"...we can conclude that individual catchments do not generally or consistently follow Budyko curve trajectories as posited by the catchment trajectory conjecture..." (page 17 line 8)

what the authors have (re)discovered is what happens when one violates a key model assumption.

Hence, their assertion that

"...this conjecture in hydrological analyses (e.g., precipitation partitioning sensitivity and causal attribution to anthropogenic and climatic impacts) will likely introduce significant errors and may lead to spurious conclusions."

seems to be difficult to support from their empirical test but also seems unfair both to the model itself and to those in the community who apply the model in accordance to its inherent limitations.

Response 2:

First, we wish to make it clear that our intention is not to be unfair to the Budyko framework, nor to those who apply it. Our conclusions about the fundamental limitations of the parametric Budyko equations emerged from a genuine interest in the framework and a careful study of the catchment hydrology literature while attempting to improve the biophysical understanding of the catchment-specific parameters. After realizing the non-transferability and under-determined nature of the parametric Budyko equations in our own applications, we decided it would be beneficial to bring these issues to the larger catchment hydrology community, particularly given the large number of recent papers using the parametric framework. We have attempted to convey a message of recontextualization of prior results in the manuscript (e.g., page 3 lines 15-17, page 21 lines 11-14). However, we can likely improve our representation of this message, as also suggested by **Reviewer 1.** We will address this by explicitly illustrating how the intent and efforts of prior work can be maintained if an appropriate interpretation of the parametric Budyko framework is applied (see proposed edits to the manuscript in our responses to **Reviewer 1 Comment 3** and **Reviewer 1 Comment 16**).

Next, we thank the reviewer for his accurate representation of the results from our empirical test of the catchment trajectory conjecture. Specifically, we found that the long-term behavior of hundreds of US and UK reference catchments generally follow the non-parametric Budyko curve when taken as an ensemble, however the behavior of individual catchments over time do not follow the conjectured parametric Budyko curves. The reviewer suggests, however, that time-series catchment data should not necessarily follow a single explicit curve, since such data violate the underlying steady state assumption of the Budyko approach. He therefore suggests that our conclusions about the catchment trajectory conjecture are not supported by our empirical test.

However, the reviewer's interpretation neglects two central elements of empirical test methodology. Specifically, we used only reference catchments (i.e., stable catchment characteristics) and accounted for potential non-steady state storage impacts by testing catchments' actual trajectories through Budyko space for essentially all possible temporal averaging windows (see further details in our response to **Comment 3**). We thus believe that we fairly tested the important assumptions and interpretations of the parametric Budyko framework—and found them to be unsupported for individual catchments. While the ensemble behavior of many catchments generally follows the non-parametric Budyko curves, our evidence suggests the trajectories of individual catchments are not specific parametric curves.

To be clear, we agree that not accounting for storage dynamics is a violation of the Budyko framework's underlying assumption. However, we note that the catchment trajectory conjecture and methods derived from it specifically suggest that the temporal evolution of a catchment (i.e., its time series) will follow a particular Budyko curve under changes in aridity index. For example, the two derived methods that the reviewer quotes from our manuscript, precipitation partitioning sensitivity and causal attribution to anthropogenic and climatic impacts, are fundamentally based on the idea of change in a catchment's Budyko space trajectory over time. Combined with the reviewer's **Comments 3 and 4** below, we thus conclude that the reviewer is suggesting that time series data will follow a particular Budyko curve, as long as the catchment properties remain unchanged and the storage dynamics are properly accounted for (or the assumption of steady state is not violated). We agree with this interpretation of required conditions (i.e., stable catchment characteristics and steady state storage) and contend that our empirical test methodology meets these requirements and therefore provides a robust assessment of the claims of the catchment trajectory conjecture.

<u>Comment 3:</u>

Is it not the case that only time-series (daily, weekly, yearly etc) hydrological data that have water storage change effects accounted for can be used to test, empirically, whether the catchment specific parameter is temporally constant?

Response 3:

Yes, in order to quantify $\overline{E}$ accurately for a catchment (which is required to test if the catchment-specific parameter is temporally constant), the remaining components of the water balance must be known since, $\overline{E} = \overline{P} - \overline{Q} - \overline{\Delta S}$. While the 728 references catchments used in our empirical analysis had daily time series of P and Q (which can be averaged over the time interval of interest to obtain $\overline{P}$ and $\overline{Q}$), they lacked estimates of ΔS, as is the case for the vast majority of available catchment hydrology data. We recognized this potential limitation in our analysis and addressed it explicitly by testing the catchment trajectory conjecture for all possible relevant realizations of each catchment's actual trajectory through Budyko space (page 11 lines 26-29 and page 12 lines 1-3). Specifically, we computed time-varying $\overline{P}$, $\overline{E_0}$, and $\overline{E}$ by applying moving-average window sizes ranging in annual steps from 1 year to the full length of record (which ranged between 12 and 45 years for all catchments).

It should be expected that above some threshold averaging window size (e.g., 10-, 20-, 30-year average window), changes in catchment storage average to zero ($\overline{\Delta S} \sim 0$). This has been shown to be the case for catchments across Earth (Han et al., 2020), with 71% of catchments reaching steady state with a 10-year averaging window and 94% reaching steady state with a 30-year averaging window. However, even if this threshold behavior does not apply universally, it should be expected that, for some averaging windows and some catchments, steady state conditions would be present (i.e., $\overline{\Delta S} \sim 0$). In either case (threshold behavior or not), testing all the relevant averaging windows for all catchments allows for a robust test of the catchment trajectory conjecture. In the case of threshold steady state behavior, if the catchment trajectory conjecture is correct, actual and expected Budyko space trajectories for a catchment would be consistently and statistically indistinguishable once the averaging window reaches a sufficient length (see page lines and page 12 lines 4-13). Statistically, this means that the frequency at which actual and expected Budyko space trajectories were found to be statistically indistinguishable would be higher than expected at random (i.e., more than 5% of all actual vs. expected trajectories would be statistically indistinguishable at a significance level of 0.05). An elevated frequency of statistical similarity would also occur if catchments were only rarely in steady state. The reason for this elevation is that the number of statistically similar trajectories from averaging windows where $\overline{\Delta S} \sim 0$ would be in addition to the number of statistically similar trajectories expected from random chance.

Critically, the results of our empirical test presented in the main text show that the catchment trajectory conjecture is not supported. Out of the 24,501 actual trajectory realizations, 23,231 (95%) were found to have consistent differences (p-value < 0.05) from their "expected" trajectories, while only 1270 (5%) were found to be statistically indistinguishable. This proportionality is exactly what would be expected due to random chance. We emphasize that our methodology directly addresses the (very common) lack of knowledge of catchment storage dynamics that the reviewer points out, and it thus provides a robust test of the catchment trajectory conjecture. However, we believe we can be clearer on this point in the manuscript, and so we suggest the following edits:

1) Change the section starting on page 11 line 26 and ending on page 12 line 7 to,

"For a given reference catchment, estimates of $\overline{P}$ and $\overline{E_0}$ were obtained from daily records of P and E_0, while estimates of $\overline{E}$ were calculated from the catchment water balance, $\overline{E} = \overline{P} - \overline{Q}$, which assumes impacts from storage dynamics are negligible ($\overline{\Delta S} \approx 0$). Since $\overline{P}$, $\overline{E_0}$ and $\overline{E}$ represent temporal averages, and we were also interested in temporal trajectories of those magnitudes, we computed time

series of moving averages for each of the three variables. The temporal averaging window for which $\overline{\Delta S} \approx 0$ is typically unknown, however, it has been shown to exhibit a threshold behavior (i.e., above a certain averaging window size $\overline{\Delta S}$ is consistently negligible) (Han et al., 2020). The threshold averaging window size can vary between catchments, though approximately 71% of global catchments reach the threshold with an averaging window of 10 years while 94% of catchments reach the threshold with an averaging window of 30 years (Han et al., 2020). To address the uncertainty in the threshold averaging window size, we computed different "realizations" of the actual trajectories in terms of $\frac{\overline{E_0}}{\bar{P}}$ and $\frac{\bar{E}}{\bar{P}}$ for each catchment for all possible integer-year averaging windows in annual steps from one year to the full length of record. Then, the theoretical (or "conjectured") Budyko curve of Eq. (5) was fitted by adjusting the value of n using the full length of record in each catchment.

The conjecture was tested for each reference catchment by statistically comparing all realizations of its actual trajectories to its theoretical Budyko curve trajectory using the non-parametric sign test (Holander and Wolfe, 1973). This is a distribution-free test for consistent over- or under-estimation between paired observations (see also Supplemental Information Sect. S2). If the catchment trajectory conjecture is correct, then the frequency at which actual and expected Budyko space trajectories are found to be statistically indistinguishable will be higher than what is expected due to random chance (see also Supplemental Information Sect. S2)."

2) Add a section to the supplementary information,

Section S2.5 Controlling for potential catchment storage dynamics

"The temporal averaging window for which $\overline{\Delta S} \approx 0$ for the reference catchments in this study is unknown and may vary between catchments. However, it should be expected that above some threshold averaging window size, $\overline{\Delta S} \approx 0$ for many of the catchments most of the time (e.g., greater than a 10 year-average window), otherwise the reference catchments would rarely be in steady state. This threshold behaviour for $\overline{\Delta S} \approx 0$ has been shown to be near universal for catchments across Earth (Han et al., 2020), however, even if this was not the case for our reference catchments, and they are rarely in steady state, it should be expected that for some averaging windows for some catchments, steady state conditions are present (i.e., $\overline{\Delta S} \approx 0$). With or without this threshold behavior, testing all of the averaging windows for all catchments allows for a robust test of the catchment trajectory conjecture. In the case of threshold behavior, once the averaging window has reached sufficient size for a particular catchment (e.g., > 10 years), if the catchment trajectory conjecture is correct, the actual and expected Budyko space trajectories would be consistently statistically indistinguishable. This means that the frequency at which actual and expected Budyko space trajectories are found to be statistically indistinguishable would be higher than what would be expected due to random chance (i.e., more than 5% of all the possible actual and expected Budyko space trajectories would be statistically indistinguishable at a significance level of 0.05). An elevated frequency of statistical similarity would also occur if catchments were only rarely in steady state. The reason for this elevation is that the number of statistically similar trajectories from averaging windows where $\overline{\Delta S} \approx 0$ would be in addition to the number of statistically similar trajectories expected from random chance."

Comment 4:

Might it be that water storage provides the link between the analytical meaning of the parameter and the physical understanding of how individual catchments have different E for the same P and Ep? That is, when a stored water term is included as a water supply term alongside P, then it introduces a catchment-specific and time-dependent term into the model fundamentals and into your analytical analysis?

Response 4:

We agree that it might be possible that explicitly including storage changes in the water supply term (i.e., $\bar{E} = [\bar{P} - \overline{\Delta S}] - \bar{Q}$) would show that reference catchments' trajectories follow a particular parameter Budyko curve with a temporally constant catchment specific parameter (a "storage-dependent catchment trajectory conjecture"). However, to our knowledge, this hypothesis has not been explicitly tested within the literature. Additionally, the results from our empirical test (which implicitly incorporates storage dynamics) do not support this idea (see response to **Comment 3** above). Therefore, while it might be possible, the best currently available evidence does not appear to support such as hypothesis.

Furthermore, if the storage-dependent catchment trajectory conjecture is indeed correct, it can only be correct for one of the formulations of the parametric Budyko equations, since the various functional forms produce contradicting curves/trajectories when the catchment-specific parameter is held constant (see Sections 3.2.2 and 4.2.2 of the main text).

References:

Han, J., Yang, Y., Roderick, M. L., McVicar, T. R., Yang, D., Zhang, S., and Beck, H. E.: Assessing the Steady-State Assumption in Water Balance Calculation Across Global Catchments, Water Resources Research, 56, 10.1029/2020wr027392, 2020.

Holander, M., and Wolfe, D. A.: Nonparametric statistical methods, New York: John Wilew and Sons Inc. Publications, 497, 1973.