Comments in black
Replies in blue

Referee #2

Gnann et al. focus on the robustness and stationarity of streamflow sensitivities to P and Ep because the concepts of sensitivity and the methods used to estimate it are not fully clear in the literature. They approach this by: (1) generating six combinations of synthetic data from the Turc-Mezentsev model to identify methods that perform reliably across conditions; and (2) applying the selected methods to catchments with long-term observations to explore how sensitivities evolve over time and the sources of uncertainty. The manuscript is well structured and several of the results are very insightful. My detailed comments are below.

We thank the reviewer for their detailed and constructive review.

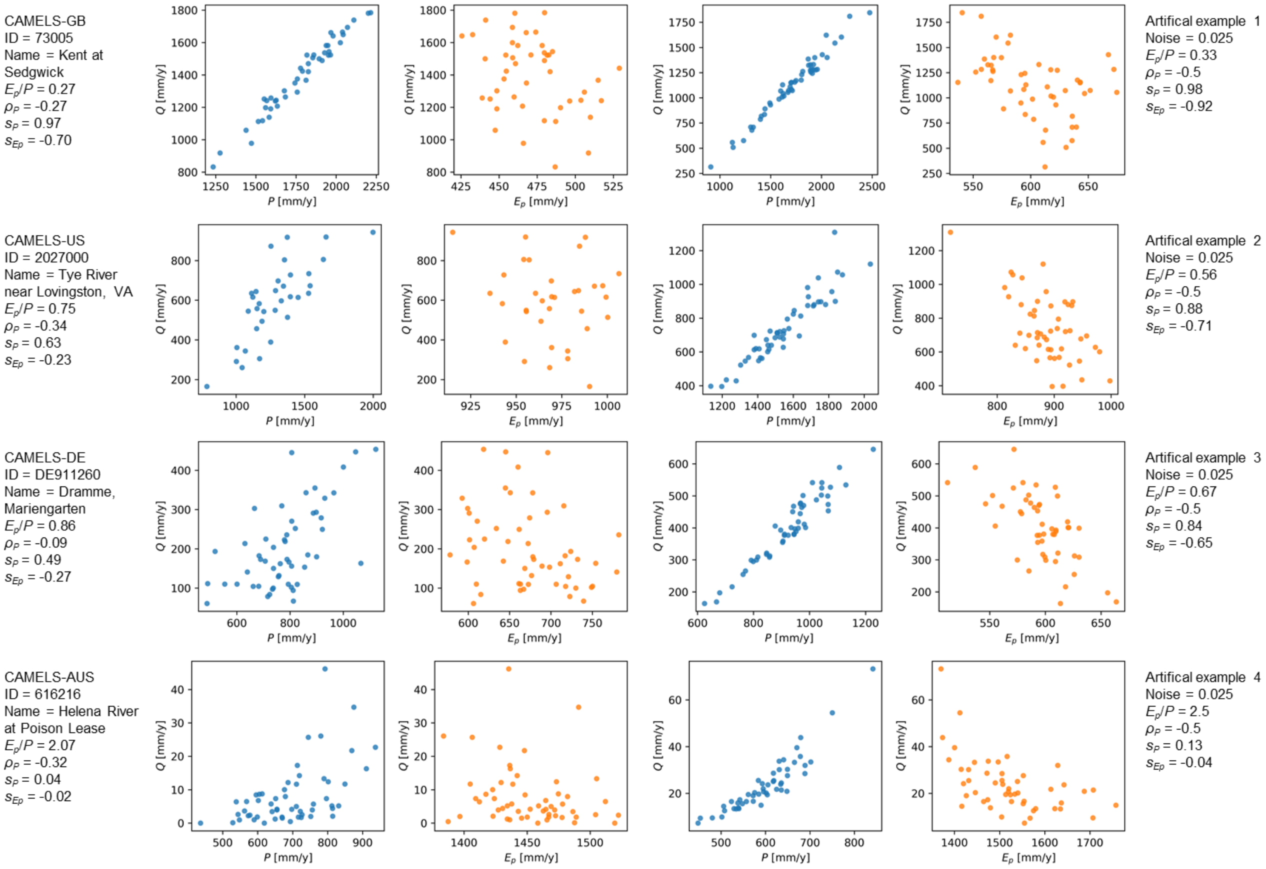
Major comments:

before going into the specific results, I think it would help if the manuscript clarified how the theoretical sensitivities relate to the empirical ones. The analytical sensitivies come directly from the Turc-Mezentsev curve, whereas the empirical values are estimated from interannual variability using regression. Because the Budyko curve is nonlinear, a regression slope over many years does not necessarily match the local derivative at the long-term mean. This difference might explain part of the mismatch between the analytical lines and the observations in several regions.

This is a good point. Indeed, the assumption behind the regression is that variability around the long-term mean is sufficiently linear. When looking at the Budyko curve we can see that this is approximately accurate if aridity only varies by a relatively small amount (e.g. 0.1 to 0.2, which for typical catchment averages may translate into a few 100 mm/y). In addition, some parts of the curve are more curved than others, likely making a linear fit less accurate.

The fact that we can recover the sensitivities very well (in absence of noise and correlation) shows that the regression approach generally works well in the range of variability used (here 15% and 5% standard deviation). We can also visually check the degree of (non-)linearity for some example cases, as shown below. We note that this bivariate view is somewhat limited, though, given that we actually look at a trivariate relationship here. We picked 4 catchments from the 4 countries investigated, ordered according to aridity, and 4 artificial cases that have a similar range in P and E_p. We can see that the pattern is linear in most cases, but that weak nonlinear behaviour is visible for the more arid catchment (both in the observational data and in the artificial data). What is also visible from the figures, is that real catchments exhibit more scatter, resulting in lower sensitivities. This is particularly evident for the German catchment and matches the general pattern found here, namely that the German catchments show lower sensitivities on average.

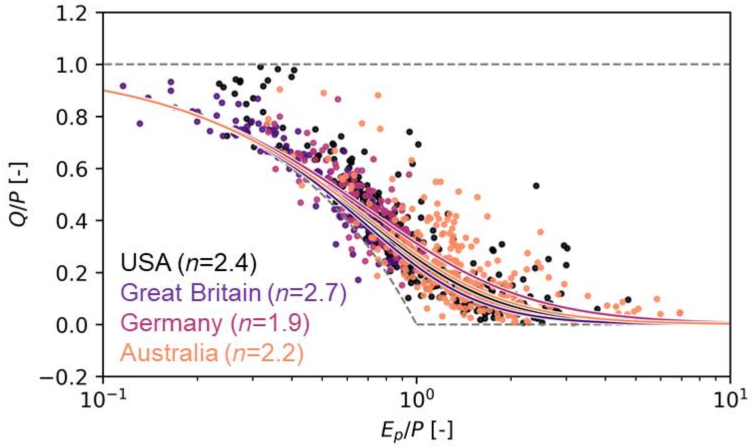
We will add the figure to the supplement and discuss this issue more thoroughly in a revised manuscript. One option for future work would be to quantify the degree of nonlinearity and to check in which catchments such behaviour most likely occurs



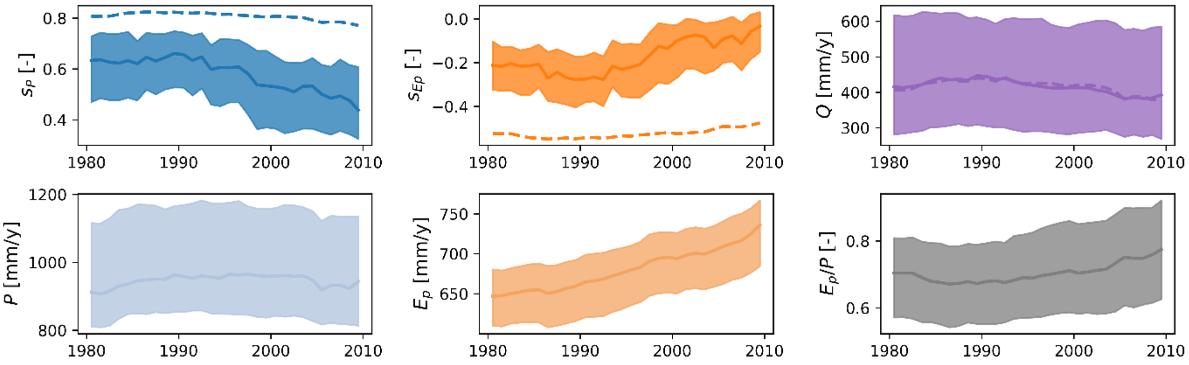
related to the point above, using a single Budyko parameter n across all catchments can also affect the comparison. Since n controls the curvature of the Turc-Mezentsev relationship, regional differences in n would naturally show up as differences in the “expected’’ sensitivities, even though the manuscript notes that the exact value of n is not the focus. A fixed n can still influence the shape of the theoretical trends, so it would help to check how sensitive the analytical results are to this choice. This may be particularly relevant for Figs. 7 & 8, where the theoretical line captures the trend over Australia but not Germany. Labeling points by country in Fig. S1 might reveal if this mismatch is regionally systematic. The strong bias in German trends also make it difficult to interpret the degree of non-stationarity, even though this general pattern is consistent. Maybe consider to use boxplots or similar summaries to describe the catchment-level trends.

Thank you for bringing this up. One main reason why we do not want to focus on the parameter *n* is that it lumps together many, often co-related factors (e.g. related to storage, vegetation, seasonality, groundwater losses/gains etc.) and may even compensate for systematic data uncertainty (e.g. E_p calculation, underestimation of *P*). So, independent of how it is assessed, interpreting the results will not be straightforward, especially not without an in-depth analysis.

We agree, however, that the mismatch can partly be attributed to the fact that the Turc-Mezentsev model does not capture local/regional conditions well. We thus calibrated the *n* parameter (by minimizing the absolute error) for each national dataset to (a) get an idea how variable it is and (b) use this later on in the country-based trend analysis. For all other analyses (e.g. Figures 5, 6), we now use a calibrated value for all catchments (2.2). The figure below shows the catchments and the fitted curves. The lowest value is found for Germany (1.9), while the highest is found for Great Britain (2.7). We will add this to the supplement.



Using these *n* values for calculating the expected trends shifts them slightly, but the general mismatch remains, as shown below for the German catchments.

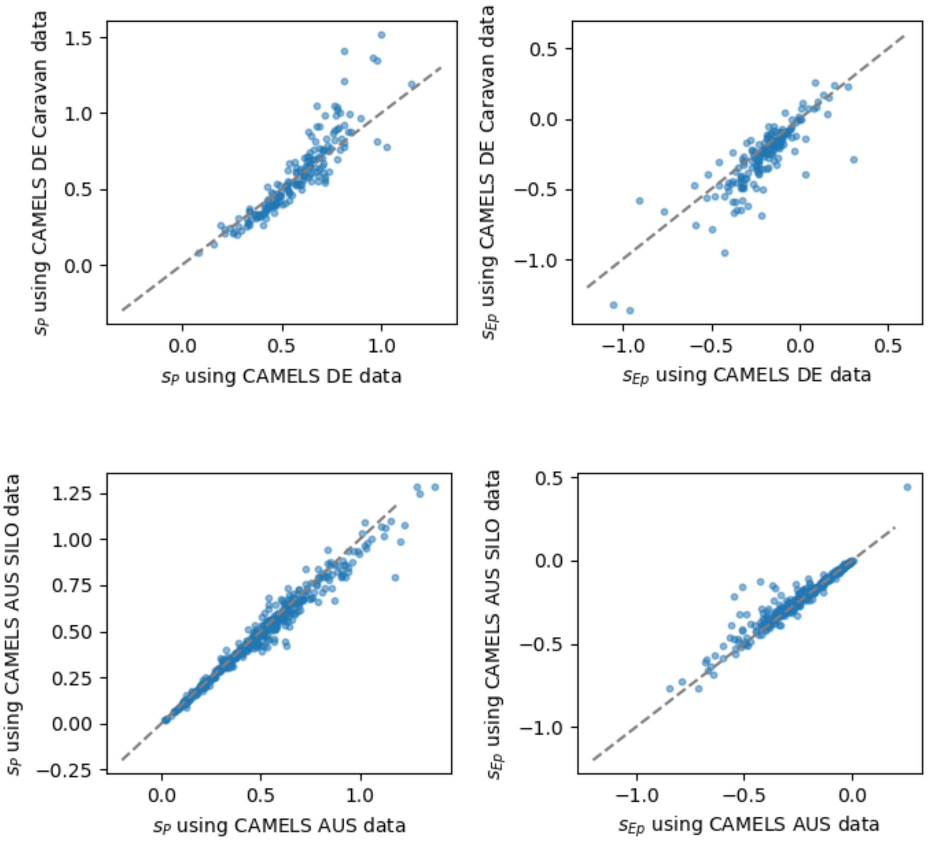


The main reason for the differences therefore seems to be a different one and is likely related to the fact that the Turc-Mezentsev does not capture the sensitivities and their temporal variability/trends well, even if it captures the mean water balance well. (As a side note, this implies that using the Turc-Mezentsev calibrated on many catchments does not predict well how individual catchments respond to changes in aridity.) We will discuss possible reasons for the differences between theoretical and empirical trends more thoroughly in a revised manuscript, see also our reply to R#1.

for the data used, the national forcing of P is very useful, but more information is needed on how Ep is calculated in each region. Why are different formulations used in these datasets? If these formulations were chosen because they best represent local conditions, it would be good to explicitly clarify. If not, part of the apparent non-stationarity in sensitivities might come from the way Ep is estimated. This might also help explain the positive s_Ep values in Fig. 4b. A brief comparison with an alternative Ep method (such as PET_Yang2019) might be benificial, although I think it may be somewhat beyond the main scope.

The reasons for using the different data products vary. Sometimes there just happens to be a national-scale product (e.g. Australia, Great Britain), sometimes the decision was made during the creation of the dataset, often as part of a modelling exercise (e.g. USA, Germany). We assume that the respective authors chose the E_p estimation methods to work well for the national conditions. For example, in the CAMELS-DE paper it reads “This variant of the Hargreaves formula resulted in the lowest mass balance error in most catchments with respect to other methods (e.g. Penman, Priestly–Taylor) to estimate evapotranspiration […].” We will add some more details on the E_p estimation methods used.

As briefly described in l.182-188, we compared sensitivities estimated using different P and E_p products for Germany and Australia, which were available in the respective datasets or as part of their Caravan extension. The scatter plots below show the resulting sensitivities using multiple regression #1. For Germany, for instance, the Spearman rank correlation between the national and the Caravan-based sensitivities are 0.94 for s_P and 0.86 for s_{Ep}, and the mean absolute differences = 0.08 and 0.09, respectively. While this shows that there are some differences, it does not explain the apparent non-stationarity, nor the positive s_{Ep} values. But we totally agree that E_p is one of the most uncertain inputs here and will further emphasize this in our revised manuscript.



for the unexplained variation in sensitivities, it might be useful to discuss the role of vegetation. The vegetation cover influences the rainfall-runoff relationship, water storage; and the effective Budyko parameter n. Long-term changes in vegetation traits could shift catchments relative to the theoretical curve and influence sensitivities to both P and Ep. Nijzink and Schymanski (2022) provide an interesting example of how adjustments in vegetation influence the Budyko parameter n, and connecting this to your results might strengthen the interpretation.

Thank you for this useful reference. We will discuss the role of vegetation and other potential influences on the sensitivities more thoroughly in a revised manuscript.

Minor comments:

for Line 114 & 117, could you provide these results in supplementary materials?

Sure, we will add a comparison of sensitivities estimated using 5-year averages and sensitivities estimated using annual averages to the supplement.

Lasso and ridge regression lead to virtually the same values, so we do not include them here.

for Table 1, (a) Log-log regression should be log-linear regression and e_{EET} should be e_Ep; (b) why do you use PET and Ep together?

Thanks for pointing that out. That was a mistake and should all read E_p. We will fix it. We will also change it to log-linear regression.

for the Turc-Mezentsev model, how is Ep calculated? It directly influence s_Ep, s_P and Ep/P.

There is no need to calculate E_p for the theoretical experiment. Both the P and E_p values are chosen as described in Section 2.3.

When I first saw Table 4, I misunderstood the relative trend, i.e. positive s_Ep with a negative relative change. I think this table is unnecessary.

We will remove the table and add the trend values to the figure caption.

Reference:

1. Yang, Y., Roderick, M. L., Zhang, S., McVicar, T. R. & Donohue, R. J. Hydrologic implications of vegetation response to elevated CO2 in climate projections. Nature Clim Change 9, 44–48 (2019).
2. Nijzink, R. C. & Schymanski, S. J. Vegetation optimality explains the convergence of catchments on the Budyko curve. Hydrology and Earth System Sciences 26, 6289–6309 (2022).