Comments in black
Replies in blue
Changed text in purple

Note that we changed some responses, which now focus on actually implemented changes.

Referee #1

SYNTHESIS

This paper deals with the precipitation and potential evaporation sensitivity of streamflow. It presents a theoretical study on the impact of different uncertainty sources which is very original, and allows to discard definitively one of the classical methods to identify elasticity (never seen anywhere in the literature... would be worth a technical note in itself). Then the paper goes on to show that the ongoing climatic change has already changed the empirical precipitation elasticity of streamflow in Germany, a very interesting and original result in itself.

We thank the reviewer for the detailed assessment and the helpful comments.

OVERALL COMMENT

This is a very good paper: excellent substance, excellent analysis, excellent form.

I would like in particular to congratulate the authors for using the sensitivities / absolute elasticities which are easily and logically interpretable (and have easily identifiable physical limits) instead of the relative ones ('true' elasticities). The plots showing the dependency of the relative elasticities (derived from the Turc-Mezentsev formula) to aridity, published elsewhere in the literature may be mathematically right but is useless in hydrological terms (the behavior with aridity makes no sense: we, as hydrologists, are not interested to know that a theoretical ratio of two terms that tend towards zero has a mathematical limit, we are interested to know that the two terms tend towards zero).

As a reviewer, my only recommendation is "don't change a word and publish as it is".

But since I am not only a reviewer but also a hydrologist interested in the topic, I could not help to comment your paper below. Feel free to consider or not my suggestions. I realize that there is enough matter to publish several very interesting papers, and I am definitely not requesting you to turn this paper into a very long undigestible paper.

Honestly, my only regret is your title, which is a little vague and not at the level of your work. The fact is that there are several very interesting points in your paper, it may be difficult to choose one over the others. Also, I guess that a strict statistician would argue that the term non-stationarity is not well-chosen, and would prefer you to talk about changing behavior, I remember a discussion with Prof. Koutsoyannis 10 years ago on this topic (see e.g. Efstratiadis et al., 2015).

Thank you for bringing this up. We updated the title which now reads: Uncertainty, temporal variability, and influencing factors of empirical streamflow sensitivities

DETAILED COMMENTS

The introduction is excellent

Thank you!

Fig1: why did you choose to plot 1-Q/P and not Q/P? Also, I have a problem with your physical limits: when you use catchment data the water limit should correspond to Q=P and the energy limit to Q=P-EP. Your water limit only corresponds to the "physical limit" (Q=0). I agree that it will be strange to have a "water limit" at 0, this is why I would personally use prefer to plot Q/P and not 1-Q/P. See e.g. Fig. 1 in the paper by Andréassian et al. (2025).

This is a good point and we changed the figure and the figure caption in our revised manuscript.
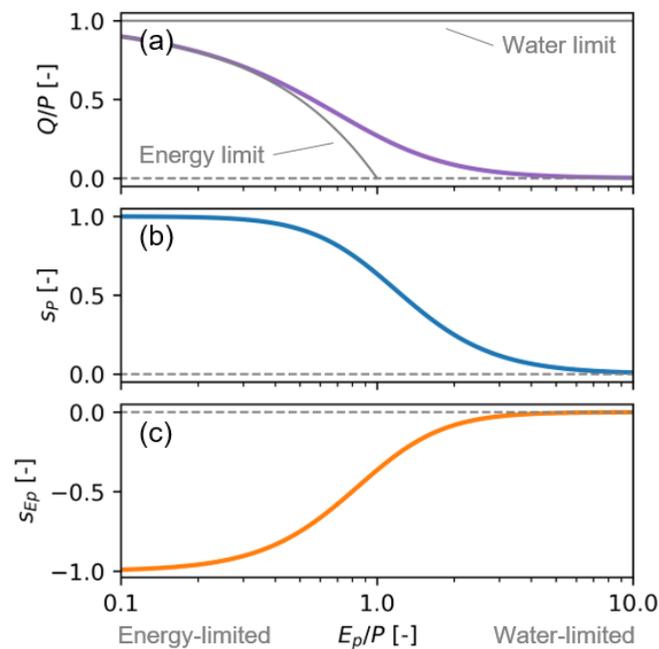

Table 1: please adjust your notations (PET -> E_P) for homogeneity with the rest of the paper

Thanks for spotting this, we fixed this.

Table 1: I understand what you mean by "same as $Q = s_P P + s_{\text{PET}} PET + c$" even if I do not completely agree: when expressed in deltas, the formula offers other opportunities, such as the pooled regression which you mention, and is interesting to reduce the uncertainties and yield hydrologically coherent values

What we meant is that for the present case, it practically makes no difference. Since the c parameter "incorporates" the means, the fitted sensitivities are exactly the same (we also tested this to make sure). We now changed the phrasing and write:

Leads to the same sensitivities as $Q = s_P P + s_{E_p} E_p + c$

Table 1: I do not understand why you introduced eq #1... a very (too?) simplistic choice, unless you want to show us that when the elasticity coefficients try to adjust to represent at the same time the intercept of the regression and the slope, strange things happen (do you really need to add this option in this paper?)

We added multiple regression #1 because it leads to the "best" sensitivity estimates in the present case. We did not originally expect that and we agree that setting the intercept to 0 is a strong assumption. But since it leads to the lowest error (Table 3) and to the most realistic range of sensitivity values for potential evaporation (Figure 4b), we decided to keep it. In addition, we think that its link to the complementary relationship is interesting as it also highlights the strong assumptions inherent in that relationship (namely that the elasticities have to sum up to 1). We had another read through our manuscript and made sure to always state that this method is by no means a perfect choice.

l130: does it make sense to assume positive correlation between delta Q et delta E? 0 and negative could be enough? Unless there are places with that type of correlation in Australia perhaps.

As can be seen from Figure S3 in the supplement, there are indeed only a few catchments with positive correlations (16 to be precise; located in the US, Great Britain, and Germany). But since there are some in most of the studied countries, we decided to keep it.

Also in the generation procedure, I note that all the catchments are considered conservative. However, in many datasets (made of mostly small catchments) catchments are significantly contributing to regional aquifers, they "leak". I am not sure it's worth to include this aspect in your theoretical experiment, but it could be worth to discuss potential specificities of "leaking" catchments in the discussion part (the fact is that the leaked quantity itself can be sensitive to Precipitation).

This is a good point. It is true that this could also be tested in our theoretical experiment and it would certainly be quite interesting to do so. One way to do so would be the inclusion of a systematic bias in addition to random error. For instance, we could reduce streamflow each year by 10% to mimic a leaky catchment. The effect should be relatively straightforward, though, at least if the reduction is 10% each year: losing catchments will show lower sensitivities (but the same elasticities) because the absolute streamflow anomalies show a smaller spread. This is similar to precipitation undercatch, where the opposite should happen, because precipitation variability will be underestimated.

So, since this happens in real catchments, it certainly also explains some of the variability compared to the Turc-Mezentsev model. From a practical point of view, the difficulty then lies in identifying places where this is actually the case, which is probably too much for the current manuscript.

We did not add any additional analyses, but discuss this possibility in our revised manuscript.

This suggests similar or larger uncertainties for observational data, where additional factors may complicate the estimation of sensitivities, such as systematic biases (e.g., due to precipitation undercatch) or water balance issues (e.g., due to inter-catchment groundwater flow). This could also be studied using a synthetic experiment, but is beyond the scope of this paper.

l170 Temporal trend: it is not very clear which type of trends you apply (I understood later that you just use the observed trends: it would be good to add a sentence here on this).

Thanks for pointing that out. We have expanded the description, which now reads:

[…] To do so, we first average P, E_p, and Q from the observational datasets (described in Section 3) over 20-year moving blocks in a longer time series (at least 50 years), resulting in at least 30 blocks. We then calculate both theoretical sensitivities and streamflow values (using the Turc-Mezentsev model forced with P and E_p) and empirical sensitivities (using multiple regression on P, E_p, and Q) for each 20-year block. Finally, we calculate trends of theoretical and empirical sensitivities over time with the Theil-Sen trend slope estimator (Sen, 1968). In order to calculate relative trends (in %), we normalize the trends by the sensitivities calculated in the first 20-year block (i.e., at the beginning of the time series). The trend analysis is carried out separately for different subsets of observational data (described in Section 3), for which we use an independently calibrated an n-value each to ensure that the Turc-Mezentsev captures regional conditions reasonably well.

Fig3: I understand now that the nonparametric method of Sankarasubramanian et al (2001) gives wrong (i.e. positive) sensitivities of streamflow to E$_P$. Did anybody in the literature ever mentioned that?

Good question. The method is mostly used for sensitivities to precipitation, since studies investigating sensitivities to potential evaporation are usually aware of the strong confounding effect of precipitation, thus using multiple regression or Budyko-type approaches. So, to our knowledge, it has not yet been reported but also not often been used. We added an additional sentence to the discussion.

The nonparametric method and single regression are therefore unreliable sensitivity estimators in most cases, especially for of s_{Ep}.

l240, Fig 4: it would be interesting to note that one of the methods respects (almost) the physical limits, while the other does not. But I still do not understand why you introduced the multiple Regression #1 with intercept set to 0. It was obviously a bad choice... the model uses the degree of freedom of s$_P$ (which is often not even statistically significant) to compensate for the lack of intercept in the regression: you end up fitting Q = aP +b and not Q = aP + bE$_P$

We already mentioned that one of the methods mostly respects the physical limits in l.236-237 of the manuscript, but now made this clearer.

Theoretically, we would expect s_P to range from 0 to 1 and s_{Ep} to range from -1 to 0. That is, s_{Ep} should always be negative, as an increase in evaporative demand should be related to a decrease in streamflow. However, we find that 52% of values for s_{Ep} are positive when using method #2, while it is only 3% when using method #1.

Regarding multiple regression #1: as discussed above, we agree that it is not a perfect choice, but we did want to present it as it yields interesting results. In addition, if we were to use the results for s_{Ep} from multiple regression #2 in the rest of the manuscript, we would end up with poorly defined sensitivities to potential evaporation. This would make the follow-up analyses much less clear.

Given that multiple regression #2 is the most commonly used method, we now added all the plots shown in the paper using multiple regression #2 instead of #1 to the supplementary information.

Fig6: very interesting graph

l265: when you look at the change of sensitivities with aridity defined as a variable it is like using a second-order assumption of "space-for-time trading". You should mention it.

We added the following sentence to our revised manuscript.

The fact that the sensitivities estimated empirically using year-to-year variability show a similar pattern as the sensitivities obtained by calculating partial derivates of a spatial relationship (Eqs. 12 and 13) hints at some degree of space-time-symmetry. However, the theoretical values tend to be underestimated and there is substantial variability, clearly more than for a Budyko-type plot of the same catchments (shown in Figure S3 in the Supplementary Information).

l274: this is a very interesting and very original result of your paper. I would be particularly interested to understand whether it is linked to a change in seasonality (not accounted for in the regressions), to the incapacity of the E_P model to represent the true evolution of the evaporative demand of the atmosphere, or to some other factor... A suggestion (for another paper) would be to test a monthly or daily rainfall-runoff model. If it is able to represent the change of precipitation sensitivity, then the problem is due to the seasonality that the annual anomaly model cannot account for. Otherwise, it means some other hydrological process is unaccounted for (or the E_P model is inappropriate).

This is an interesting question. As noted in our initial response, when plotting the evolution of seasonality over time (using the seasonality index based on Woods, 2009), we find little overall changes in Australia and a slight increase in Germany, suggesting that rainfall has become slightly more summer dominant. This, however, does not really explain any of the trends.

We now decided to explore the role of seasonality by splitting up the Australian dataset into two subsets. We also discuss in more detail the potential reasons for the observed (non-matching) patterns. Please refer to our response below, in which we discuss this new analysis that we added.

Table 4: I am not sure to understand the sign of the relative values (and the necessity for them, if it is complicated to understand). Is this table really useful? Fig 7 and 8 are already extremely clear.

We wanted to give some actual numbers and not just figures, but since R#2 also commented on this table, we now moved it to the Supplementary Information. With regard to the sign: the sensitivities decrease, hence the negative sign. Since all our values decrease, we now removed the minus sign but clarified this by writing, for instance:

"between 15% and 70% decrease"

Figure 7 and Figure 8 are extremely clear and interesting. Without requesting too much additional analysis (because it would turn your paper into a book...), I was wondering whether the Australian dataset would allow for producing 2 subsets one with P and E_P out of phase (the Mediterranean part of Australia), and another with P and E_P in phase: I believe this could help to interpret the trend observed in Germany. Another solution would be to compute an index characterizing the P-E_P phase-shift.

Thank you for the suggestion. We now split up the Australian dataset into two subsets using a seasonality index P_S. In addition, we also had a look at the distribution of the baseflow index (BFI) for each subset to aid interpretation of the results. The distribution of the seasonality index and the BFI are shown below, as well as the trends for the three subsets.
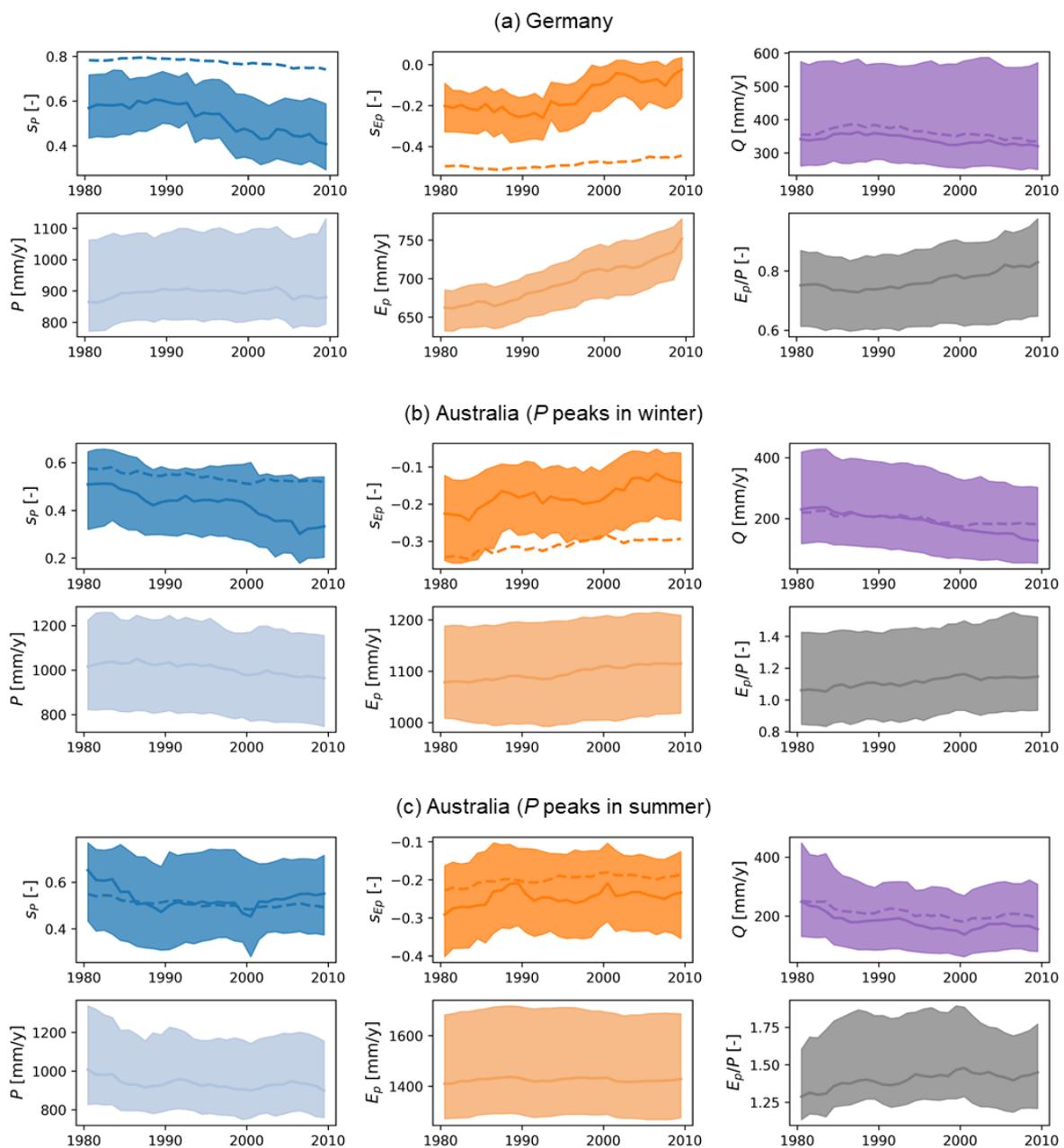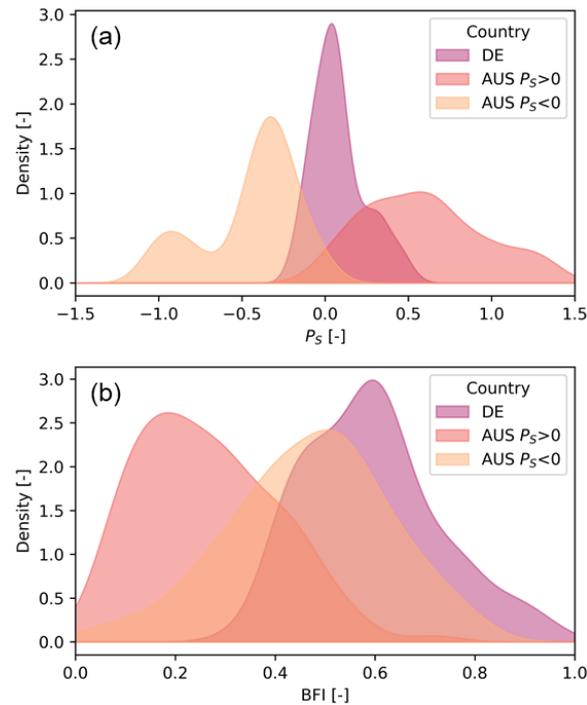
In brief, we found that catchments with different seasonal distributions of precipitation also show different patterns, but that the Australian catchments where precipitation peaks in summer, which are somewhat closer in seasonality to the German catchments, actually are more different in terms of overall behaviour. There are multiple hypotheses that could explain this behaviour, one of them being that storage (as indicated by the BFI, cf. BFI distributions) is relatively more important for the German and the winter dominated Australian catchments.

The results are also discussed in the text.

Especially for the German catchments and the Australian subset where most precipitation falls in winter (Figure 7), the theoretical and empirical trends show large differences. For both subsets, we find that the decreasing trend is larger than predicted by the analytical model and that the sensitivities themselves are lower than the Turc-Mezentsev estimates. It is unlikely that these differences are solely due to data uncertainty, because they generally remain when different data sources are used (cf. Section 3). Rather, the results suggest that these catchments (or at least their year-to-year variations) are less driven by annual totals of precipitation and potential evaporation, for instance due to larger influence of catchment storage, precipitation seasonality, or changes in land cover and human impacts. Interestingly, the average trend in streamflow is roughly captured by Turc-Mezentsev model calibrated to each subset, but the sensitivities are not (see Figure 7). This substantiates that year-to-year variability is, in many catchments, not necessarily a good proxy for responses to longer term climatic changes.

The influence of precipitation seasonality is not conclusive, as the pattern found for the German catchments where most precipitation falls in summer is more similar to the pattern found for the Australian subset where most precipitation falls in winter. Yet the German catchments and the Australian subset with winter dominant precipitation both have relatively high BFI values on average, much higher than the Australian subset with summer dominant

precipitation (see Figure S18). In line with the previous finding that catchments with high BFI have lower sensitivities, the same catchments might also respond less directly to climatic trends, for instance due to longer term groundwater responses. Finally, while we fitted the n-value to each subset for the trend analysis, using a fixed value might also explain some of the difference between analytical and empirical trends, as it has been shown that n might change as a consequence of vegetation adapting to new conditions (Nijzink & Schymanski, 2022). These present just some hypotheses that might explain (parts of) the trends, but that require more extensive testing.

l336: you write "More than half of the values based on method #2 are larger than zero". But I guess that anyway the p-values from a Student t-test would consider these values as non-significantly different from 0. Perhaps mention it?

When using method #2, indeed many of the p-values for s_{Ep} are smaller than 0.05 (if we chose this as a typical, but arbitrary, threshold). Interestingly, though, the values that are significant are both positive and negative. For method #1, many more values are significant, though of course that might in part be because the second term compensates for a lack of an intercept. We now added the p-values to the supplement and briefly discuss them in the text.

The p-values, shown in Figure S10 in the Supplementary Information, are almost always larger than 0.05 for s_P (in more than 99% of the catchments). The p-values for s_E vary, with values being smaller than 0.05 in 83% of the catchments for method #1 and only in 16% of the catchments for method #2.
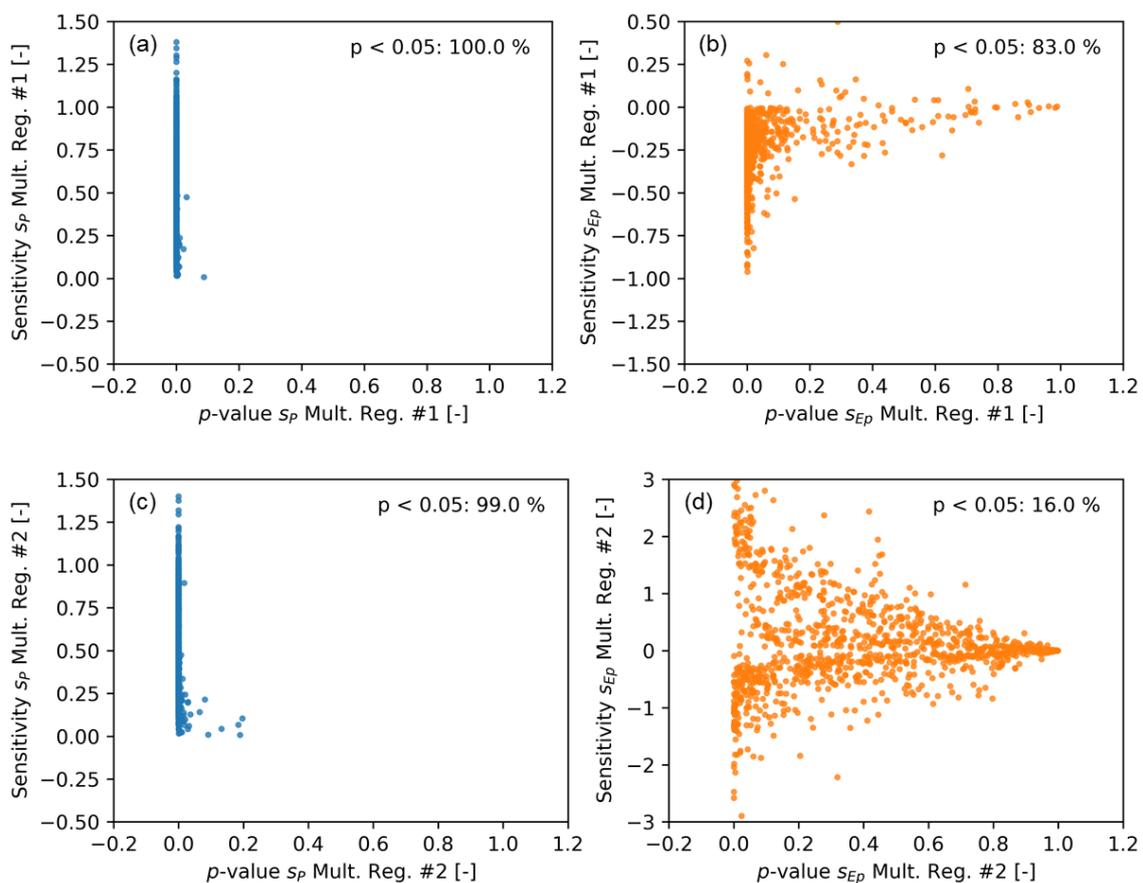
REFERENCES

Andréassian, V., Guimarães, G.M., de Lavenne, A., and Lerat, J.: Time shift between precipitation and evaporation has more impact on annual streamflow variability than the elasticity of potential evaporation, Hydrol. Earth Syst. Sci., 29, 5477–5491, https://doi.org/10.5194/hess-29-5477-2025, 2025.

Efstratiadis, A., Nalbantis, I., and Koutsoyiannis, D., 2015. Hydrological modelling of temporally-varying catchments: facets of change and the value of information. Hydrological Sciences Journal, 60 (7–8). doi:10.1080/02626667.2014.982123

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. We visually checked the degree of (non-)linearity for some example cases, as shown below. 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). We note that this bivariate view is somewhat limited, though, given that we actually look at a trivariate relationship here. We added the figure to the supplement.

A related question is to what extent year-to-year variability in real catchments can be used to approximate long-term changes in the water balance. Interesting in this respect is Figure 5, as it shows that the sensitivities estimated empirically using year-to-year variability show a similar pattern as the sensitivities obtained by calculating partial derivates of a spatial relationship (Equations 12 and 13). This (as pointed out by R#1) hints at some degree of space-time-symmetry. However, there is still substantial variability around the theoretical curve. In part this might relate to the fact that year-to-year variability is influenced by other factors, such as storage. There is some evidence for this, because catchments with high BFI fall further away from the curve. We incorporated this into our revised discussion and emphasized this a bit more.

We now discuss this issue more thoroughly in our manuscript, the revised paragraph is copied below.

Equation 11 is usually applied at long time scales, for which changes in storage are negligible (so that $P = Q + E_a$). Accordingly, also the sensitivities (Eqs. 12 and 13) should be seen as sensitivities to long-term changes P and E_p. Here, we relax this assumption and use these equations to study year-to-year variability. We note that the same assumption is (implicitly) made in many sensitivity studies, which use year-to-year variability to estimate how

catchments respond to long-term change, even though there is evidence that year-to-year variability in Q also responds to changes in, for instance, storage (Tang et al., 2020). For the theoretical analysis carried out here this does not matter, as we test how well different methods can estimate sensitivities from synthetically generated data with the assumption that storage changes are zero. It will, however, matter when interpreting sensitivity estimates based on observational data. Here, a mismatch in time scales (i.e., annual variations used to calculate empirical sensitivities vs. spatial variations of long-term water balances between catchments underlying the Turc-Mezentsev model) might lead to differences between the empirical and theoretical sensitivities.

In addition, since the Turc-Mezentsev model is nonlinear, a linear regression might not match well the local derivative at the long-term mean, especially if year-to-year variability is large. To get an idea regarding how nonlinear the relationship could be for typical ranges of variability, we had a closer look at some example catchments, both synthetic ones and real ones (shown in Figures S16 and S17 in the Supplementary Information). The bivariate relationships between Q and P and E_p, respectively, appear to be mostly linear, though some weak nonlinearity can be seen especially for strongly water-limited catchments.
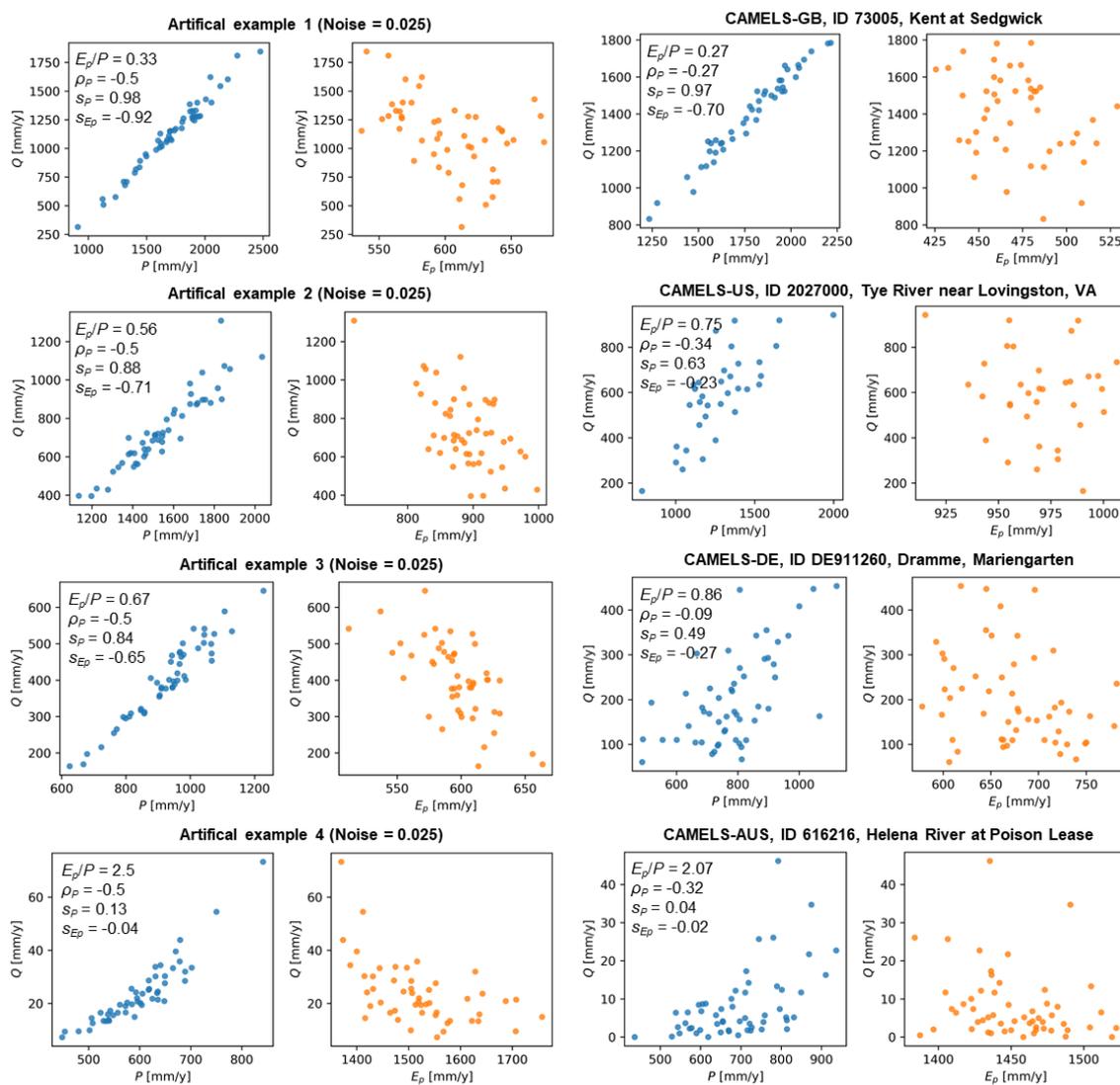
related to the point above, using a single Budyko parameter n across all catchments can also affect the comparison. Since n controls the curvature of the Turc-Mezentsev relationship, regional differences in n would naturally show up as differences in the "expected'' sensitivities, even though the manuscript 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 added this to the supplement.
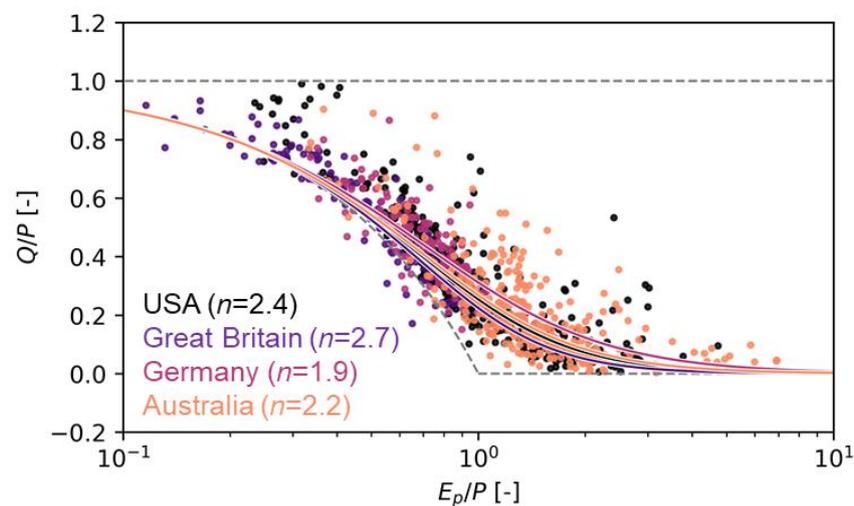
Using these *n* values for calculating the expected trends shifts them slightly, but the general mismatch remains, as shown below for the German catchments.
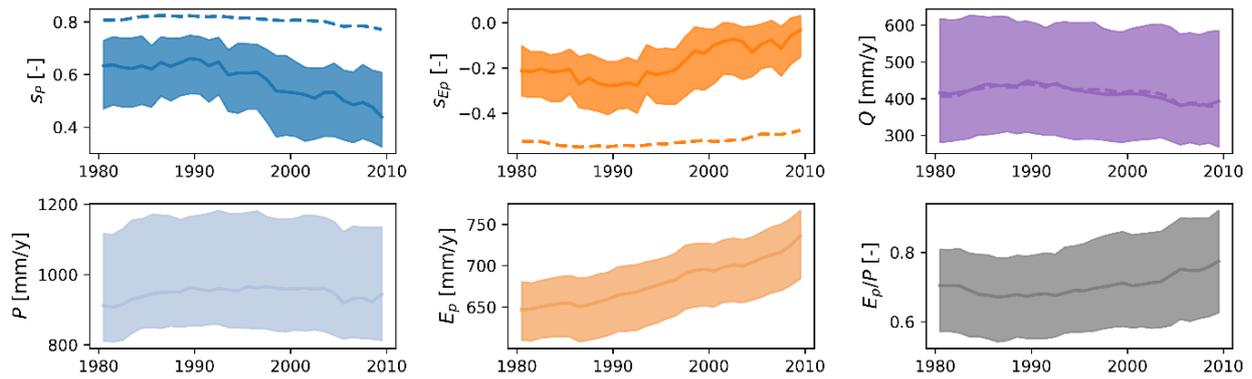
The main reason for the differences therefore appears to be a different one and is likely related to the fact that the Turc-Mezentsev does not capture the sensitivities and their temporal variability/trends well, even if it captures the mean water balance well. We now discuss possible reasons for the differences between theoretical and empirical trends more thoroughly, 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 their regional 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 now edited the section to be clearer about the E_p estimation methods used and extended the comparison and calculated the sensitivities for CAMELS-DE and CAMELS-AUS both using national forcing products and their Caravan extensions. The results, which we added to the supplement, are shown below and the text now reads:

[…] We therefore made a brief comparison to study the influence of different P and E_p datasets on the resulting sensitivities for CAMELS-AUS and CAMELS-DE, the two datasets used for the trend analysis. In particular, we estimated the sensitivities using P and E_p from national data sources (our default option here; see Table 2) and from their Caravan extension (Kratzert et al., 2023), which is based on ERA5-land and uses the Penman-Monteith equation in its updated version to calculate E_p (Kratzert et al., 2025).

Overall, the estimated sensitivities are similar (see also Figure S2 in the Supplementary Information). For CAMELS-AUS v2, we obtain a Spearman rank correlation of $\rho_S = 0.97$ and a mean absolute error of MAE = 0.09 for streamflow sensitivity to precipitation, and $\rho_S = 0.94$ and MAE = 0.07 for streamflow sensitivity to potential evaporation. For CAMELS-DE,

we obtain $\rho_S = 0.94$ and MAE = 0.08 for streamflow sensitivity to precipitation, and $\rho_S = 0.86$ and MAE = 0.09 for streamflow sensitivity to potential evaporation. In conclusion, we find that the resulting sensitivities somewhat depend on the forcing products used, but that the differences rather small and also do not substantially alter the trend analysis (not shown here). We will come back to the issue of reliably estimating E_p in the discussion, but note that this is not our main focus here.
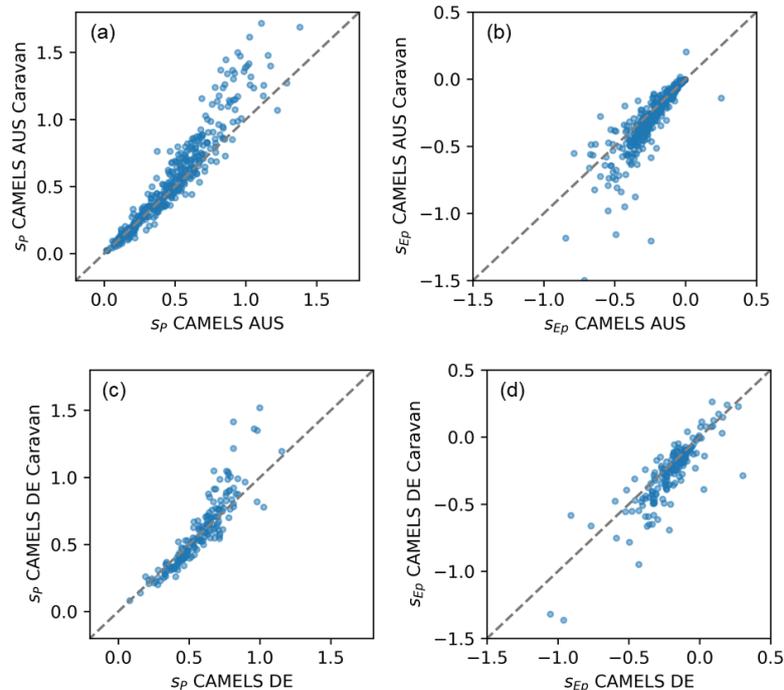
for the unexplained variation in sensitivities, it might be useful to discuss the role of vegetation. The vegetation cover influences the rainfall-runoff relationship, water storage; and the effective Budyko parameter n. Long-term changes in vegetation traits could shift catchments relative to the theoretical 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 now discuss the role of vegetation and other potential influences on the sensitivities more thoroughly in the discussion section.

Minor comments:

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

We slightly rephrased the respective lines and have now added a comparison of sensitivities estimated using 5-year averages and sensitivities estimated using annual averages to the supplement (see Figure S1). 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 fixed it. We also changed 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 now moved the table to the Supporting Information and slightly rephrased the text.

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).