Review of Catchment-scale groundwater recharge and vegetation water use efficiency, by Troch et al.

This paper presents how the Horton Index can be used to estimate groundwater recharge and groundwater dynamics. Using 247 MOPEX catchments it is shown that:

- maximum total and deep catchment storage are mainly positively correlated, and (b)
- the Horton Index can be used to predict low flows (FDC' tails) and groundwater recharge

Better understanding the link between climate, vegetation, and catchment water balance dynamics is a relevant topic, potentially suitable for publication in HESS.

However, I have a number of concerns that need to be addressed before I can recommend publication of this work in HESS, Namely:

Does the Horton Index actually indicate vegetation water use efficiency?
The Horton index has been used in several past studies, but I am unsure if it really can/should be interpreted as catchment-scale vegetation water use efficiency. The paper states that ET dominates total terrestrial evaporation, but most estimates indicate that ET is ~60% of global terrestrial E (thus ~40% has little to do with transpiration). Second, the HI definition is very closely related to a catchment's water balance; a ratio of V/W seems almost equivalent to using a ratio of E/P because E ≈ P- Q, and P ≈ P-Qd (because quick flow Qd is only a small fraction of total P for many MOPEX catchments). This tight relationship between E/P and HI (and AET/P) is also shown in the inset panel in figure 2 of this manuscript. I guess the question that I struggle to answer myself is: how can I be confident that HI actually informs about water use efficiency, and are results attributed to water use efficiency not largely the result of correlations with known controls on the water balance such as climate aridity?

The link between groundwater recharge/baseflow and the Horton Index seems trivial thus are reported correlation really meaningful? In line 98-188 the Horton index is outlined, and its link with expected GW recharge is clarified. While I found this useful to read, it also appears to say that the definition of the Horton Index is directly connected to baseflow (see lines 104-133, and later stated in the discussion "the HI is estimated from baseflow separation and is then used to estimate average baseflow conditions". The argumentation that the statistical relationships shown in the study are not the result of spurious correlation is not convincing (to me). It would be helpful if you actually line out the "several arguments that go against this statement". In addition, the used argument that "we use predicted HI from climate and landscape properties to estimate average baseflow and long-term recharge." is not a clearly explained argument, but a statement without explanation (but when I try to formulate an argument using this statement myself I end up in a similar circular problem as you stated yourself earlier in this paragraph).

Why does it matter that maximum total and deep catchment storage are mainly positively correlated? This finding is stated as one of the main results of the paper, but it is not clear to me what we learn from this relationship?

This study relies on linear fits for baseflow recession (and the inference of storage changes), but is this assumption realistic? Past studies that characterized baseflow recession using MOPEX data indicate that most catchments operate non-linearly. For example, Ye, Sheng, et al. "Regionalization of subsurface stormflow parameters of hydrologic models: Derivation from regional analysis of streamflow recession curves." Journal of hydrology 519 (2014): 670-682. Is assuming a linear relationship problematic for estimating storage dynamics or is this still accurate under non-linear conditions?

The comparison with USGS data is not convincing. First, it is stated that the method for the estimates rely of different principles, which seems only partly true. The HI may not explicitly have baseflow index in its definition, however, it's definition is obviously strongly correlated with baseflow (as the paper already acknowledges). Therefore, I cannot really see this comparison with the USGS data as a validation using independent data. Second, I agree that regional differences in GW recharge are mostly explained (i.e. R^2 = 0.77). However, predictions at individual sites seem to diverge strongly from the USGS estimates. Thus, maybe site predictions of the HI baseflow are not so good. It would be fair to point this out better.

The comparison with Q50 is unclear. It is assumed that long-term average baseflow conditions can be made from the 50th percentile of the flow duration curve (Q50), "*but that this choice was more an intuitive guess than an informed decision*". If this is all context given about the GW recharge estimate, I have no idea what this comparison means (because I have not idea what the Q50 estimate means).

Total storage is a misleading term. This paper uses "total storage" to represent something that is more often referred to as "dynamic storage" (e.g. see Kirchner 2009, WRR). Total storage seems misleading to me, because total GW storage in the landscape is much larger. Something like **"**total dynamic storage" would be more useful.

The study relies on the assumption that baseflow is a good indicator if streamflow originates from GW. While the event water vs groundwater separation is a fixture in many hydrology textbooks, we also know that taking a closer look (using isotopic data) often tends to indicate that this assumption is not accurate. What are the implications for your work?

Detailed comments

L13-18: Does HESS actually require key points? Or can these be removed?

L23-24: "can either runoff as infiltration/saturation excess or quick subsurface flow" It seems like something went awry with this sentence.

L32: I think it would be more insightful to describe the method rather than the source when you state "from the US Geological Survey".

L50: "affect" seems more appropriate than "define".

L51: Budyko elucidated the role of "climate aridity" on the long-term water balance, not the other dimensions of the climate. Therefore use "climate aridity" and not "climate".

Line 55-56: "thoroughly elucidated" suggests (to me, and thus maybe other readers), that the role of these factors is clear. However, without going into detail of fully reviewing all the factors that you mentioned, I find it hard to agree with that interpretation. Would it help to slightly rewrite this sentence to get a more nuanced perspective?

L54: does Scanlon et al, actually study the long-term WB, or solely GW recharge rates?

L56: "first-order controls of the inter-annual […] water balance" have been extensively studied. To just say "are less understood" seems too simple. Previous works need to be acknowledged, and a clearer knowledge gap needs to be identified. For example:

Potter, N. J., & Zhang, L. (2009). Interannual variability of catchment water balance in Australia. Journal of Hydrology, 369(1-2), 120-129.

Carmona, A. M., Sivapalan, M., Yaeger, M. A., & Poveda, G. (2014). Regional patterns of interannual variability of catchment water balances across the continental US: A Budyko framework. Water Resources Research, 50(12), 9177-9193.

Sankarasubramanian, A., & Vogel, R. M. (2002). Annual hydroclimatology of the United States. Water Resources Research, 38(6), 19-1.

L59-60: "several studies have investigated the role of climate, vegetation and other catchment properties in hydrological partitioning" is unclear to me. What is your definition of "hydrological partitioning" that warrants only citing this limited amount of studies. Aren't there literally thousands of studies that focus on how water is partitioned once it hits the landsurface? (probably I just misunderstand the point here).

L66: "short" instead of "small".

L88: "baseflow" or "baseflow that sustain riparian vegetation"?

L94-96: I am unsure this assumption is really justified. Maxwell and Condon report 62 ± 12% ET. Jasechko et al. 2013 reports high ET ratios, but there are well known issues with this analysis: (see: Coenders-Gerrits, A. M. J., Van der Ent, R. J., Bogaard, T. A., Wang-Erlandsson, L., Hrachowitz, M., & Savenije, H. H. G. (2014). Uncertainties in transpiration estimates. Nature, 506(7487), E1.), whereby studies provide substantially lower ET ratios (see for example: Sutanto, S. J., van den Hurk, B., Dirmeyer, P. A., Seneviratne, S. I., Röckmann, T., Trenberth, K. E., Blyth, E. M., Wenninger, J., and Hoffmann, G.: HESS Opinions "A perspective on isotope versus non-isotope approaches to determine the contribution of transpiration to total evaporation", Hydrol. Earth Syst. Sci., 18, 2815-2827, https://doi.org/10.5194/hess-18-2815-2014, 2014. And Schlesinger, W. H., & Jasechko, S. (2014). Transpiration in the global water cycle. Agricultural and Forest Meteorology, 189, 115-117.)

L227: How variable where the determined k-values between years, within individual catchments? Were those variations somewhat realistic (since one can expect them to be fairly constant, right?).
Line 235: What makes them independent? It seems that both are derived from streamflow observations, and not from independent sources (such as actual well measurements)?

Line 269: why a "linear" fit?

Line 285: why a “linear” correlation?

L293: is there actually any evidence that pumping is a big issue in this catchment, or can it just be data uncertainty?
Line 340-341: Is HI a better explanator of this than solely using the aridity index?

Line 357: “strongly” seems redundant.

Line 350-360: Explain to the reader where this Q50 baseflow estimation originate from. Without that context, it is hard to judge this comparison.

Line 444-446: “Surprisingly, the magnitude of the correlation between these variables and different storages across the region of study is very similar and only ~10% of the catchments seem to be out from these observed patterns” I am unsure what “seems to be out” really means here.

Figure 4: without more description, I find it difficult to follow the logic of this analysis from the figure.

Figure F6A2: it this really the best linear fit you get for the presented data?

Figure 7: it would be helpful to also display the non-significant locations (for example in light grey).

Figure 7: does the Horton index provide a better estimate to predict FDC;s than solely using climate aridity?

Figure 8: does the Horton index provide a better estimate to predict FDC;s than solely using climate aridity?

Figure 10: is it possible to make the comparison between the two recharge rates visually more intuitive?

Line 740: “Study” doesn’t need to be capitalized.