I have read "Estimating epikarst water storage by time-lapse surface to depth gravity measurements," by Champollion et al. The paper presents a relatively straightforward analysis of a unique dataset. I commend the authors for collecting an impressive dataset. Is it available? I urge the authors to release it online and/or state how it can be obtained. I do wonder if we are reaching the point where this type of investigation (repeat microgravity) is no longer novel and some unique aspect is necessary to justify publication. In that regard, the comparison of measurement methods was interesting (long vs short), but only minimally investigated. A side-by-side comparison at the same time would have been quite useful.

The English language usage is generally understandable but the manuscript would be improved by careful editing. In particular, it is too wordy in spots.

The introduction states that epikarst is important, that geophysics can measure hydrologic processes, and that gravity is a hydrogeophysical tool, but I don't think it sufficiently ties them together. I think it could be improved by focusing on a specific aspect of the karst-gravity relation and/or more directly stating what information gravity provides that can't be obtained with other methods.

The literature review is quite minimal, which is okay, but it gives the impression that Jacob, 2009, is the only reference for subsurface and karst gravity measurements (I agree it is the best reference for such work). There are other examples in the literature, e.g.,

> Tanaka, Y., Miyajima, R., Asai, H., Horiuchi, Y., Kumada, K., Asai, Y., & Ishii, H. (2013). Hydrological gravity response detection using a gPhone below- And aboveground. *Earth, Planets and Space*, *65*(2011), 59–66. https://doi.org/10.5047/eps.2012.06.012
>
> Koth, K.R., and Long, A.J., 2012, Microgravity methods for characterization of groundwater-storage changes and aquifer properties in the karstic Madison aquifer in the Black Hills of South Dakota, 2009–12: U.S. Geological Survey Scientific Investigations Report 2012–5158, 22 p.
>
> Kennedy, J. R., Koth, K. R., & Carruth, R. L. (2015). Surface and Subsurface Microgravity Data in the Vicinity of Sanford Underground Research Facility, Lead, South Dakota. *U.S. Geological Survey Open-File Report 2015–1043, 32 P., http://dx.doi.org/10.3133/ofr20151043*.
>
> Van Camp, M., Viron, O., Pajot-métivier, G., Casenave, F., Watlet, A., Dassargues, A., & Vanclooster, M. (2016). Direct measurement of evapotranspiration from a forest using a superconducting gravimeter, 225–231. https://doi.org/10.1002/2016GL070534.
>
> Van Camp, M., Meus, P., Quinif, Y., Kaufman, O., Van Ruymbeke, M., Vandiepenbeeck, M., & Camelbeeck, T. (2008). Karst Aquifer Investigation Using Absolute Gravity. *EOS*, *87*(30). https://doi.org/10.1029/2004JB003497.
>
> Deville, S., Jacob, T., Ch, J., & Champollion, C. (2013). On the impact of topography and building mask on time varying gravity due to local hydrology. *Geophysical Journal International*, *192*, 82–93. https://doi.org/10.1093/gji/ggs007

(Actually, there are fewer references than I had thought. That may be worth noting, and would make the release of the author's dataset that much more valuable).

The authors may wish to note other advantages of underground measurements, e.g., common-mode rejection (density-changes above and below a certain elevation will be nearly equal at all stations), likely reduced seismic noise, the ability to locate stations on bedrock, others? And disadvantages: potential for rough handling during transport, others? Any comment on the paper of Reudink et al. which suggests the meter used in the study has high sensitivity to tilting during transport?

> Reudink, R., Klees, R., Francis, O., Kusche, J., Schlesinger, R., Shabanloui, A., … Timmen, L. (2014). High tilt susceptibility of the Scintrex CG-5 relative gravimeters. *Journal of Geodesy*, *88*, 617–622. https://doi.org/10.1007/s00190-014-0705-0

69: Better to present as a circular area

70-72: "As surface gravity…" This sentence doesn't make sense to me. I think further elaboration/clarification of gravimeter sensitivity of above-ground and below-ground measurements, including discussion of common-mode rejection, is warranted. Also, The concept that storage increases above a gravity-station cause a decrease in gravity.

164: How representative are these gages? There must be much more spatial uncertainty than measurement error, suggest focusing on the former.

Can you state the similarities and differences between the two study areas?

205: You need to describe the CG-5 first, specifically, that it is used to measure the relative difference in gravity between two locations. You should mention that all of the gravity values in the paper have large (~980000 mGal) offsets relative to the absolute-gravity value. In some places, it's CG5, in others it's CG-5.

227: "Stable during the studied period." – does that mean the continually increasing calibration factor for this meter in Jacob (2009) from 2006-2009 is no longer observed?

238: The concepts of "loops" should be introduced. Were loops observed in order from the surface to depth? If so, is there concern about unequal transport time? The discussion of the "short" and "long" strategies later would also benefit by framing it in terms of loops.

258: Please check denominator in equation 2. Isn't the number of gravity readings, m, be larger than the number of gravity stations, n? That would make the denominator negative.

Where any observations removed during the adjustment? Was the Chi-square goodness-of-fit test used?

Figure 4: Please clarify that these are the residuals of the observed gravity differences vs. the adjusted gravity differences. The labels ("#93 data") are unclear –they are the number of observations, I assume? Suggest "n = 93". Maybe mention the panels which show "short-time" vs "long-time strategy"?

274: To be more precise, I think you'd say error from the relaxation must be accounted for in the error budget (you could potentially not correct for it, but still have a clear error budget).

280: I don't understand "4 and 5 in our case." This means the reference station was observed 4 times, and the other stations 5 times, during each survey? Or 4 loops were observed during some surveys, and 5 during other surveys? It is also confusing in Annex 1. Suggest replacing "4 to 5" and "1 to 2" there with "short" and "long."

I appreciate the authors testing different measurement strategies. It is unfortunate they weren't both tested during the same campaigns.

Could the authors clarify how they end up with, e.g., 140, 248, and 209 observations (for site BESS, long strategy), for a survey with 5 stations? They are including every sample at every station, instead of averaging the samples for any particular station setup? I think this leads to artificially increasing the degrees of freedom (denominator in Equation 2) and underestimating the uncertainty. In other words, there are really only 4 independent measurements (Base-1, Base-2, Base-3, Base-4), not 140+.

289: One disadvantage of the long strategy, as I understand it, is that with a single loop drift can't be separated from possible tares (offsets).

297: I would guess the better residuals with the long-term strategy is primarily a consequence of having fewer observations to adjust. If the observations at each station were averaged prior to the network adjustment, with a single loop and allowing for drift, all of the observations could be matched exactly during the adjustment and the residuals would be zero.

309: If it's known there was a gravity jump (you can know this because there were multiple loops?), why not correct for that?

321: Annex 1 presents a summary of the adjustments, not "all raw gravity data." The row labels in Annex 1 aren't visible.

340: I think it's sufficient to state "terrain, gradients, latitude, and depth aren't changing over time" and jump straight to equation 6. Equations 3 and 4 already excludes Earth tide and ocean loading effects that have been accounted for (with zero uncertainty).

370-372: "Moreover, the large scale heterogeneities…" this statement is unclear.

374-375: Neither the capacitive function nor fast transfer have been explained.

365, 378: times not time's

387-388: I'm unclear what storage variations are discussed here: variations at the land surface, above the elevation of the surface station? It would make sense to discuss these in terms of admittance factors ("the admittance factor due to rainfall stored at the surface is xx µGal/mm, at 12 m depth it is xx µGal/mm, etc.)

404: 23.8 mm of water = 1 µGal is approximately the Bouguer slab approximation (41.9 µGal = 1 m of water). As stated on line 394, shouldn't this be twice the Bouguer attraction (23.8 mm = 2 µGal)? It looks like 23.8 was used to convert between µGal and mm in Table 1. Unless I misunderstand, this is a major error that needs corrected, and impacts all of the conclusions of the paper (I'll assume I'm mistaken w.r.t the rest of my comments).

410: optimum = minimum and maximum?

411: "without": here, and elsewhere, I think the authors overstate the accuracy of the results. It may be appropriate to state the yearly cycle is measured with minimal ambiguity, but probably not without any ambiguity (uncertainty would be a better word).

411-412: Is it necessary to add all depths for the BESS site? Why not just difference the surface station and the deepest station?

415: It would be helpful to indicate, here in the text (e.g., "spring to summer") and/or in the table, which are the discharge periods.

Table 1: The reference to recharge and discharge periods doesn't seem consistent with the dates of the surveys: Feb-Aug is a discharge period at SEOU, but Nov-Sep is a discharge period at BEAU? I would like to see some discussion as to how this inconsistency affects the results, particularly the statement "Contrary to the two first sties, cumulative precipitation have a similar values…"

442: NWI should be described before Table 1 is introduced. It looks like NWI uncertainty was calculated as the root mean square of the P and AET uncertainty? Please state that's the case. Please double check, the value for BEAU Sep07-Feb08 (and maybe others) is off: sqr(17^2 + 17^2) = 24.

473: The gravity depth-profiles have opposite shapes, not nearly the same shapes.

478: Is 2.5-3.5 µGal from the error budget, or is it from the network adjustment? I think it's the latter.

Figure 6: Suggest showing this in units of mm of water, rather than µGal.

481: Shouldn't it be -1 and 2, instead of 1 and 2?

495: "Gravity measurements must be very useful…" The English is incorrect. You could say "…measurements may be useful…", or could be useful. Rarely is the word "very" useful in scientific writing.

497: The Deville paper appears to be quite relevant to the present work, yet isn't mentioned at all in the introduction/literature review.

498: Figure 6 suggests the error budget is 50 percent or larger than the gravity variation amplitude.

499: As mentioned earlier, I think you overstate the value of your data here. If the gravity differences are only due to hydrology, why are the error bars in fig. 6 so large, relative to the signal?

501: You probably mean minimized, not optimized.

712: A paper with a 2012 date probably isn't "in press". (it looks like the paper has been published).

504: If you only do one loop, how could you know if the drift is linear?

507: I don't understand the phrase "the coherence of the gravity measurements with respect to depth…"

522+: This discussion is hard to follow because it's in units of mm of water, whereas the results are presented in microGal.

533: What are 'spoiled structures'?

538: This paragraph is important, because you start to show how the gravimetric method can be useful for investigating pollution infiltration. But it's vague. Can you better describe how exactly the present data, or a different dataset, can be used in the context of pollution? I'm not a karst hydrogeologist but I'm guessing the conclusion that storage changes are greatest near the suface isn't a major breakthrough.

562+: This paragraph makes an important point, and demonstrates the major advantage of gravity, with its large region of sensitivity, vs. other methods.

587: I don't understand the reference to mudstone, a fine-grained, non-limestone sedimentary rock. Apparently packstone is a limestone with a high clay content, but I don't think the term is widely used.

Are there any references, or laboratory analyses, that can support the conclusion that the BESS/SEOU sites are characterized by fine-grained limestone?

577+: This paragraph is interesting, but largely conjecture (lots of “coulds”). I would be interested in reading about whether additional data (including microgravity – perhaps with more stations at shallower depth?) could help resolve some of the ambiguity.

607: “Water transfer properties” isn’t clearly defined. You mean properties of the aquifer, e.g., hydraulic conductivity and specific yield? Or magnitude of storage changes?

628: I don’t understand the basis for the 3.5 and 13 months figures. 3.5 refers to BESS/SEOU, and 13 to BEAU? Where do these numbers come from?

638: I don’t think you investigated heterogeneity; you used a method that averages heterogeneity.

648: I don’t see how the phrase after “Therefore” follows from the previous sentence.

659: Some additional description of gravity-differences vs. absolute-gravity measurements in the introduction or methods would be useful.

666: what is “total water stock”?