We would like to thank the constructive comments, please find our replies below. Our replies are marked in red, our corrections in the manuscript text are in green.

Reviewer 2

Basically, I find the paper interesting and also agree with the conclusions. There are a couple of good ideas, but also a few technical aspects that should be critically discussed in the revised version of the manuscript. I also do not entirely agree with the strict separation between simple balance models and complex process-based hydrological models. And that the process-based model is the better choice if sufficient data is available. It always depends... If the components of a simple water balance have been robustly determined (which can often be a lot of work and requires a lot of data) this is not per se less good than a complex process-based hydrological model (with its very own weaknesses and uncertainties such as non-unique solutions etc.).

Thank you for your comments.

Yes, we agree that many of these arguments need to be evaluated critically in a case-per-case basis.

We would not draw the line between water balance and process-based methods, instead between data-driven and process-based (because methodologically more complex non-linear methods could be applied within the presented approach, such as machine learning methods).

In our argumentation we would like to differentiate between 'observed data' and 'process data/knowledge'. In the presented case 'observed data' was easily obtainable and available, while knowledge about the hydrological system/process understanding was more limited. We found this a typical scenario of cases that are recently becoming of interest. Under such conditions data-driven techniques could rely solely on the available data.

We would also like to add, that modeling the system dynamics not necessarily requires the estimation of all water balance components. The presented linear regression approach only used the statistical relation between the climate forcing data and the lake levels. A process-based model in the same example would required knowledge about the geology of the catchment (K-values, hydraulic gradient) to make any estimations.

We would also like to emphasize, that our methodology, and the downward model development in general increases the model complexity gradually, which eventually lead into the transition to process-based models, without these being the necessarily better modelling end points. In the presented case, our findings are used in the development of a groundwater model of the same catchment. Hence we would rather argue that water balance and data driven approaches are ideal starting steps, and great supporting tools within a larger modeling campaign.

"The downward modeling can also fit organically into the development of process-based models, as will be shown."

Line-specific comments:

Line 13: Why „indirectly“?

We mean here the climate-lake level relation. In this formulation, groundwater trends, changing ecosystems and water use are more directly exposed to climatic variations, which then translated to lake level change.

Line 38: Here you start to explain what you want to do in this study. I would rather put this at the end of the introduction.

We have chosen to keep this section here as it gives an outline of the subsequent sections on the increasingly complex modeling methods, and it also serves as a good introduction to the downward model development.

Line 38: Citation style… only the year should be in brackets. This also applies to various other places in the manuscript.

This is to be adjusted in the final manuscript.

Line 47: Not clear to me what is meant by “higher resolution models of the catchment”

Corrected as:

“higher temporal resolution, or spatially distributed models of the catchment.”

Line 51-52: I do not agree with this statement. These handful of hydrological variables/flows are often quite difficult to determine/estimate. Therefore, I would not consider this a typical application in data-scarce regions. As I understand it, process-based models are more likely to be used in data-scarce areas to bridge the gaps of data scarcity (e.g. using standard parameter sets or those from comparable catchments + meteorological forcing data, which is usually quite accessible) – I am not very strict with my opinion here - it should rather be understood as a counter-argument.

Thank you for this comment.

We would still argue that while process-based models are more “realistic” in theory but still approximate. In practice, the data requirements to run, parameterize and test them are so high that we have large uncertainties (that we often don’t quantify). In such situations the theoretical superiority can be overwhelmed by increased uncertainty. This applies even more so for data-scarce regions; of we are honest in these applications then we would have very wide prior parameter distributions etc. for these applications. On the other hand, data-based models often work satisfactorily, especially in forecasting, but maybe for the “wrong” reasons. So best, arguably, is a simplified process-based model that relies on effective parameters estimated from data, traditionally called “conceptual model” in hydrology (despite the ambiguous meaning of the term).

In the revised manuscript, we have replaced the term data-scarce with limited-knowledge, as we find it more fitting based on your comment.

We have also added a critical sentence on our methodology from this perspective to the conclusions:

“Hence in data-scarce regions, robust process-based approaches might be a better solution as they are capable of transferring knowledge from other comparable catchments, although without data they would operate with large uncertainties.”

Line 100: Should be “used”.

Corrected.

Line 102: What do you mean by "efficient"? I find the word somewhat unsuitable here and would rather write something like "easy to use".

Revised accordingly.

Line 121: Maybe add something like “cannot yet be set up with the required level of confidence”.

Revised accordingly.

Line 136: “physics” is a somewhat controversial term in this regard, I would just write “process-based”.

Revised accordingly.

Line 149 and line 152: What do you mean with “we propose”? If this is part of your study, just delete “propose”. If you propose this for future research (based on the findings of your study), it rather belongs to a conclusion section.

Corrected.

Line 162: Why have you used ERA5 data and not data from the DWD weather station in Potsdam. This data should at least be compared to each other (for validation of the ERA5 dataset). Yes, ERA5 provides actET (as various other products as well), however, how reliable is that? In my experience, not very reliable… At best, this should be compared with the nearest station data (lysimeter) or at least discussed critically (perhaps this is not so relevant because the statistical analysis does not require very precise data… anyway requires some discussion for my taste).

We do not use the ERA5 data directly, but a dynamically downscaled dataset of it (CER v2) specifically created for the Berlin-Brandenburg region. The CER v2 dataset has a higher spatial resolution (2 km spacing) and was validated against 211 DWD weather stations, including the Potsdam DWD station.

We have run all our scripts using weather data from the Potsdam station, but could not achieve the same model fit quality as with the CER data.

The CER v2s dataset was especially better during extreme rainfall events, where the spatial variation in the amount of rainfall can be very large. Because summer storms have very significant effects on the lake levels, it was not possible to close the water balance using DWD station data only.

We were able to recreate the actual evapotranspiration dynamics using DWD station data as well. Here the biggest difference was that we were able to consider the actual land cover composition over the catchment using the gridded CER v2 dataset.

Please see our edits:

L161 “As its predecessor CER v1, the CER v2 dataset has been produced by an observation-based model approach. Global ERA5 reanalysis data has been dynamically downscaled using the Weather Research & Forecasting (WRF) model, and validated against 211 weather stations.”

And

L169 “There are two advantages of using a dynamically downscaled gridded dataset instead of relying on interpolated station data. First, such an approach provides an estimate of actual evapotranspiration for each grid point using land cover, vegetation and soil data, and dynamic data on soil moisture, while station-based observations are typically restricted to potential evapotranspiration (lysimeters or eddy flux towers would be available at only very few locations). Second, this approach explicitly takes account for meso-scale heterogeneity of weather systems, which is of particular importance for precipitation and actual evapotranspiration with high variability at spatial scales of a few kilometers or less."

Line 183: Which catchment area you are talking about? Relevant would be the groundwater catchment (right?), however, I worry that the surface catchment area is meant here. In the case of a relatively flat relief (and low hydraulic gradients) and also the fairly strong influence of groundwater extraction, it is quite likely that the subsurface and surface catchment areas differ.

This is correct. The subsurface catchment could also change in time, as we state later in the conclusions. Due to the lack of knowledge, we had to rely on the surface catchment area when defining our models, but we make this a point of discussion to learn about catchment processes in our case.

Note that the differences in catchments are already included in our models implicitly with the net subsurface inflow term, which sums up all the subsurface in and outflows from the catchment area.

Our assumption is that the dynamic changes that control the lake levels are happening within the defined catchment. The results support this assumption, otherwise the model fits should be significantly worse.

We revised the text as:

"Precipitation ($P_{catchment}$) and actual evapotranspiration ($ET_{A,catchment}$) over the (subsurface) catchment area, and not just the lake, strongly influence subsurface flow processes that feed the lake."

Line 201: "lake level changes are linearly related to storage changes" In general, I think this assumption is far too simplistic. I don't think this oversimplification is necessary either, as only a bathymetric model would be needed to represent this correctly. A quick Google search shows that such a bathymetric model exists and is accessible.

This is true if we consider the storage in the lake only, but more realistic considering the whole catchment for which bathymetric models are not sufficient. Bathymetric maps are available for Sacrower lake but not for the GGS.

Line 209: The application of the strong oversimplification, mentioned in the previous comment, quite likely has a systematic impact/error on the estimation of dF and thus the identification of tipping points (e.g., presumed tipping points might actually be related to lake bed morphology).

Interesting, thank you. We added this argument as a possible explanation for the tipping points in our discussion section.

With regard to the two previous comments, you could possibly also argue that in your particular case the water level decrease only takes place in a relatively small range and therefore such a linear assumption is not completely wrong... (but should be checked with the bathymetric model)

We think the linear assumption is reliable for the catchment scale water balance model.

"Focusing on water level changes in a limited range the linear assumption should be a good approximation, but note that under complex hydrological conditions this relation may be different. Still, this would be seen if the observed and modelled timeseries would not come close during this optimization."

Line 228: Delete the word "same" or replace it by "respective".

Revised accordingly.

Line 239: I not really get why filtering helps neural networks? Such a statement should be explained.

Removing some of the higher-frequencies, remove some of the non-linearity of the problem. This makes the calibration of high-parameter models less ill-posed, hence more robust.

In this paper filtering was necessary for visualization, hence we modified this section to avoid any confusion as:

"Optionally, input data might be filtered prior to the analysis, which would help with some more complex data driven modeling methods, such as neural networks. In this study we used Butterworth filters from the scipy.signal python package. For the autocorrelation analysis in section 4.1 a bandstop filter is used, that removes the 365 days period signal from the lake level data. For the plots of the linear regression analysis (Figure 6,7,8 and 9) a lowpass filter was used over the lake level data, with a cutoff frequency at 20 days to help with the visualization of the analysis."

Line 278: What exactly do you mean with "civil engineering information"? Please, specify.

Revised as:

"and limited information civil engineering informationon the rainwater infrastructure (e.g. manhole cover locations) is available."

Line 300-306: This belongs to the results section.

We moved these lines to the results together with lines 285-290 to the results.

Line 308-310: These lines are superficial.

We deleted these lines.

Line 324: Isn't that actually a pretty vague result? Couldn't it be that the memory size might not be larger, e.g. 100 days? How would the model results change if a different memory size were assumed?

Figure 3b shows that a very wide range of memory sizes are suitable for modeling. We have checked, and the modelled results of the linear model are very similar, even when using an unoptimal memory of 150-200 days. This is due to the fact that the model coefficients of higher days are small compared to the days closer to t=0 (see Fig. 7).

We have noted this unsensitivity in the revised manuscript:

"The large range of optimal fits indicates the robustness and insensitivity of the linear regression method."

Line 325: "One might link this time to the catchment size, as the distance travelled by the groundwater flow." I would also delete this sentence. (It's not a good style to assume that that the reader doesn't know the very obvious fact that this is caused by pressure transmission…)

Revised accordingly.

Line 415-427 and Figure 8: How representative are the selected time periods in 2006 and 2016? Is it valid to draw general conclusions from them (they could simply be singular phenomena…)? Either this should be well argued in the text or average values could simply be used as in Figure 9.

Although we present only single events, we have looked through the whole timeperiod to draw our conclusions (and we found these examples to be representative). These two periods were good examples, where the differences in the model behaviors can be discussed in detail. We found this more hands on, than only showing the averaging of Fig. 9.

Averaging could not show the impact of single rainfall events, or drought periods, as these events happen different times in different years, hence we find it important to show this also.

We expanded as:

"In this plot we zoom into two different parts of the dataset to compare the two models directly in detail, focusing on typical weather events."

Line 444: Instead of "since 2015", you should provide the exact time period, i.e. "between 2015 and XXX". This also applies to later cases in the text, e.g. line 451.

We have checked and corrected all instances to "between 2015 and 2022".

Line 461: Why only 4%? Even if one takes 50% as the average (which could definitely be regarded as the upper limit), the influence would only be marginal.

4 % was taken from an online draft version of the paper (Schleich and Hillenbrand 2009), but the actual published number is 7% (as an upper bound).

We have corrected it and added the reference to the revised version.

Line 487: "some of the relevant information is only available on one side of the lake" Would be good to be a bit more specific, i.e. which information is missing on which side.

This refers to the Study site and data section where:

L276 "From the Berlin side, data regarding water supply, wastewater management and canalization maps are available on the city webpage for multiple years. From the Brandenburg side, geological and limited information civil engineering information the rainwater infrastructure (e.g. manhole cover locations) is available."

Figure 10: Why you took 2010 as a baseline? And is a baseline necessary? For Figure 10b you could also just sum up the anomalies (e.g. as deviation from the mean) starting at zero. And why are the cumulative anomalies going down when the NDVI increases (and are above the reference/baseline value)? – if I misunderstood something, it should be better explained…

We choose a baseline for visualization purposes. 2010 was chosen as it is at the middle of the timeframe, and its value is close to the mean NDVI of the 2002-2015 period. The mean of the complete timeseries would consider the period after the investigated tipping point, which we wanted to avoid here.

The NDVI anomaly was defined as ($NDVI_{2010}$-NDVI), so larger NDVI makes the anomaly go down to show the increase in water uptake. We have added this definition to the figure caption.

"In Fig. 10b we calculated the cumulative sum of the NDVI anomaly relative to the 2002-2015 period average, and obtained a similar trend to the water balance anomaly. This suggests a possible connection between the two trends."

Line 504-505: Basically, you are saying that an increase of 10% in NDVI causes an increase of 5-15% in ET. Do you have any reference for this or at least indications for this assumption?

Unfortunately, there is no easy way to link NDVI to ET without using vegetation models. To show an estimate we looked directly at MODIS satellite data from the catchment which showed this increase in ET.

We have rephrased as:

"Beside the observed 10% increase in NDVI, MODIS evapotranspiration data, shows 5-15% increase in forest evapotranspiration."

Line 508: One of the "However" is too many.

Corrected.

Line 512-521: I agree to the hypothesis that the evolution of groundwater levels at a regional scale has an impact (although it doesn't explain the tipping point). However, doesn't that somewhat contradict your model assumption that your subsurface catchment corresponds to your surface catchment?

Please see our earlier response to the catchment definition again.

Regional groundwater trends could impact the subsurface inflow to the catchment, which could lead to lake level loss. The tipping point can happen if the connection between the lake and the groundwater is lost.

"(Lischeid, 2021) analyzed lake and groundwater level timeseries in the region with principal component analysis. The author's concluded that lakes situated on the higher parts of this lowland region are more sensitive to falling water levels than lakes in the valley bottoms because lakes situated higher are prone to losing their direct connection with the groundwater."

The groundwater flow directions could also change, that could lead to a tipping point, similarly as discussed for the vegetation in L568:

“Another explanation is that the increase in vegetation on the west side of the catchment reduced the groundwater levels locally so that it altered the groundwater flow regime. The gradient of the groundwater table in this area is very small ($3 \times 10^{-4}\ m/m$), hence a local decrease in recharge could divert the groundwater flow and modify the subsurface catchment size.”

Line 540-541: Good point! But, again, it somehow contradicts your model assumption as described in the previous comment.

See our previous reply.

Line 546-570: This is rather a summary of your results than a conclusion of your study. I would more focus on pros and cons of your study (maybe also some of my comments/critics can be discussed here).

Thank you. We have extended our conclusions in the revised manuscript according to your suggestion. We have also trimmed down some of the summary to further streamline the text (deleted lines 557-561).

“This set of methods provided an effective toolset for understanding lake level changes and their drivers in a case, where prior hydrological system and process knowledge was limited.

The developed water balance and data-driven models provided very good fits with lake level observation, which shows not just the potential of the modeling approaches, but also the applicability of the CER v2 weather dataset. The approach revealed the main drivers of the lake level dynamics, and provided some insight to systemic changes in the hydrological system, which led to some possible hypotheses regarding the lake level loss.

The presented methodology however was not able to clearly identify the exact reason behind the non-climatic lake level loss, and the proposed hypotheses can only be proved or disproved with additional experiments and/or process-based modeling.

Another drawback of the presented methodology is the strong reliance on good quality data. Closing the water balance, or obtaining a good fit with the linear model was possible only, because of the high accuracy of the weather dataset. Due to the spatial variability of precipitation, replacing it with weather station data would lead to a significant drop in model accuracy. Hence in data-scarce regions, robust process-based approaches might be a better solution as they are capable of transferring knowledge from other comparable catchments, although without data they would operate with large uncertainties. ”

References

Schleich, Joachim; Hillenbrand, Thomas (2009): Determinants of residential water demand in Germany. In *Ecological Economics* 68 (6), pp. 1756–1769. DOI: 10.1016/j.ecolecon.2008.11.012.