Reviewer 1

The authors combine the PIRFICT approach to transfer-function noise modelling with ECWMF ensemble forecasts to predict groundwater levels about a month ahead with daily time steps. They compare the results with more naïve prediction methods and show the added value of Ensemble Groundwater Prediction (EGP) over e.g. using climatology for fast reacting systems without snow dynamics. Their results are not that surprising as similar conclusions have been drawn for streamflow already. Nevertheless, the paper provides one of the first attempts of setting up an EGP system and is worth publishing after revisions.

We thank the reviewer for his/her constructive comments on our manuscript. We agree that the results are not entirely surprising, as similar results have been obtained for streamflow. As mentioned, this has not been documented too often for groundwater. Furthermore, we hope that the results from our study also motivate others to further explore ensemble forecasting in groundwater, both in research and in practice.

Two more major issues.

I miss the comparison with open loop simulations. With groundwater, we often encounter wells where we had observations in the past (so a PASTAS model can be fitted to these) but then the well was terminated. For these locations, ESP could still be done, but without updating. I suggest to add two additional experiments where seasonal forecasts are generated without updating at the 6 wells while comparing EGP with using the climatology. This would indicate how much is lost if persistence cannot be exploited compared to updating and also what the added value is of ESP over climatology in these situations (which could be considerable, also in slow systems).

Thank you for this comment. It seems only one open-loop experiment is outlined in the comment, but perhaps we misunderstood (in which case, please let us know). We have actually done the experiments without the updating part, but thought it would be too much to also add this to the paper. Considering the reviewers' comments, however, we are happy to add the results from these experiments to the manuscript. This shows the importance of updating for groundwater level forecasting, particularly in slowly responding systems. This also has practical implications, because it confirms that obtaining near real-time groundwater level data is beneficial to improve the forecast quality. We will add the results of the EGP and climatology without updating, and add this to the discussion and results as well.

I find the lack of bias-correction problematic, and the reasons stated for not doing this in lines 417-418 entirely unconvincing. A simple bias adjustment using a drizzle correction for P and a month-specific bias-adjustment (additive for T and multiplicative for P) is easy and can also be done on the fly. Also, if needed, a moving window adjustment of correction factors can be applied if the ECWMF ensemble forecasts improves over time. I bet results will improve even more if this is done.

We understand this comment from the reviewer, and want to further explain our decision to not bias-correct the meteorological forecast data:

- We want to split the bias-correction of the meteorological forecast data from the methodology of the impact model. For operationalization, MeteoSwiss will provide bias-corrected forecasts. This is a common approach for many impact models developed in Switzerland where the bias-correction is split from the development of the impact model.
- Bias-corrected forecasts from MeteoSwiss have been available only since 2018, which would drastically reduce the period available to test the EGP model.
- The pre-processing (bias correction of meteo temperature and precipitation) shows the highest positive effect in mountainous regions with dominating snowmelt processes (e.g., https://doi.org/10.5194/hess-23-493-2019). Except for Davos, all the groundwater stations are located in the lower lands, where the pre-processing showed only minor improvements compared to post-processing of hydrological forecasts (https://doi.org/10.1175/JHM-D-21-0020.1).

We will better explain in the manuscript why no bias-correction is implemented, also in the introduction of the manuscript.

Minor issues

Why is a forecasting horizon of 32 the target? Where does it come from?

This is driven by the lead time from the ECMWF forecasts. In the future, this will be extended to a 46 days lead time.

Figure 2 does not have simulated levels as stated in line 230.

We will change the sentence and remove “simulated”.

Equation 1: did you check if the noise at the six locations behaved as a first-order autoregressive process (or discrete version of the Ornstein-Uhlenbeck process)?

We checked this assumption, and for the majority of the wells, this is valid. We will adapt the manuscript to discuss this point, and the implications for forecasts at the wells where this is not valid. This is indeed important, and a point for potential improvement of the EGP model.

Line 158: please provide this formula, which, I believe in your case, with the same value of the noise variance per ensemble member, simply means adding the result of Equation (3) to the Variance of the 5100 ensemble members.

We will add the formula to the manuscript, this is indeed the results of eq 3 to the variance of all the ensemble members.

In the recommendations for future research, it would be good to also consider what needs to be done to provide space-time ensemble predictions. For instance, if one needs to assess which fraction of a region has groundwater levels below a given threshold. This can either be done with many wells (how many?) or in a spatio-temporal framework (groundwater flow models or regionalized time series models). Any thoughts on this matter?

This is an interesting point that is also part of active discussion in Switzerland, where the idea of “representative wells” is also used. We will add our thoughts on this to the manuscript and provide recommendations towards providing space-time groundwater level predictions.