Reviewer 2

This manuscript describes an ensemble groundwater forecasting method that is applied to produce daily groundwater level forecasts to lead times of 32 days. The method uses a conceptual model to simulate groundwater levels in response to climate forcing and generates forecasts groundwater level forecasts using downscaled, but not bias-corrected, extended-range forecast forcing from ECMWF. A simple auto-regressive error correction model is applied to update forecasts using recent groundwater level observations. Forecast uncertainties are derived from two sources (i) ensemble forecast forcing and (ii) an ensemble of parameters for the conceptual model. The method is applied to six groundwater wells with a range of characteristics that allow the exploration of sources of forecast skill, and the performance of forecasts is assessed using several standard forecast verification metrics. This is a nice study, and one of only a handful of investigations into forecasting groundwater. I have a couple of more major concerns and some minor suggestions.

We thank the reviewer for his/her constructive comments to our manuscript, particularly regarding the estimation of the prediction intervals (PI). We will test the proposed method to estimate the PI to investigate if this improves the estimation of the prediction intervals and may solve the issues with underconfidence of the forecasts.

Section 2.3 introduces a residual error model (equation 1) that has two terms an autoregressive term and a white noise term and the authors argue this introduces only one additional parameter. There appears to be a parameter that has not been considered here, i.e. the variance of the white noise term, which is unlikely to be equal to 1 given the range of predictions. This white noise term appears to be subsequently neglected but is likely to form an important source of predictive uncertainty of the combined groundwater level and error correction models.

Thank you for this comment. The noise term $\upsilon(t)$ is actually not a parameter but a random variable that is the result of applying the AR1 noise model (also called an error model in other contexts). This is why the error model only adds one additional parameter. There is a direct relation between the variance of the noise and residual terms through the AR1 model, outlined in eq. 3. That forecast variance increases from the variance of the noise at h=1 to the variance of the residuals as the lead time increases (as h > infinity). We will rewrite the text in lines 140-155 to better explain this part of the methodology, also in response to the comments of the other reviewers.

Section 2.5 then introduces a 3-step process for ensemble groundwater prediction. The methods third step (line 148) appears to be generating a parametric estimate of the forecast error from 5100 individual ensemble members, using the autoregressive parameter from the noise model, and using the parametric estimate of forecast error to obtain prediction intervals. I think there are two improvements that that could be made to this third step that may improve the performance of the verification metrics. Firstly, the prediction intervals could be computed directly from the ensemble members rather than from a parametric estimate of the forecast error. This would circumvent the need to assume the ensemble follows a normal distribution and therefore potentially address the limitation identified in the discussion related to the heteroscedasticity of forecast errors.

The third step is done for each ensemble member using the autoregressive parameter from the noise model. We will test the proposed method, computing the prediction intervals directly from the ensemble, and check if this improves the metrics. Will clarify in the discussion that the remark about the heteroscedasticity is related to the model residuals, and not the forecast errors, which we think is not clear from the manuscript now.

Secondly, the estimation of the forecast error uses the autoregressive parameter derived from the residual error model introduced in Section 2.3. The forecast errors are unlikely to have the same autocorrelation structure as the simulation residuals of the conceptual groundwater model because the characteristics of the rainfall forcing, observed rainfall in the case of the residual error model and forecast rainfall in the case of the forecast errors, will be different. Rather than seeking to describe the forecast errors using a parametric distribution, this third step could simply add noise reflecting the hydrological model simulation errors following the autoregressive updating -

which is likely to increase the spread of the forecast ensemble and therefore address the issue of underconfident forecasts that appear to exist for many wells (Figure 7).

We thank the reviewer for this comment, which we find very interesting. We assume in this study that the errors of the forecasts also behave as an AR1 process, but indeed did not check this. We will try the proposed method to see if the overconfidence is reduced in the proposed way and report the results.

Practical significance of the work... I would like to see a bit more context on groundwater dynamics and management in Switzerland added that would highlight the significance of the work for practical applications. Personally, I think of groundwater levels as a variable that changes relatively slowly and therefore a forecast with 30 days lead time is unlikely to have practical benefit for groundwater management decision-making. However, it is plausible that some groundwater systems may respond over 30-day time horizons and therefore the forecasts for these periods could potentially be practically useful.

This is an important remark that we will address by adding a few more sentences to the manuscript, particularly the introduction, explaining the practical significance of the work. In many locations in Switzerland the groundwater levels respond rather quickly (i.e., droughts can develop over weeks), particularly in the alluvial valleys. In other locations the response is much slower (i.e., droughts develop over months). Depending on the monitoring well, the value of the forecasts may thus be different. More generally, there is interest from governmental agencies in groundwater forecasting for droughts, particularly for the development of the Swiss drought platform. The long-term goal is to extend the forecast period to 46 days, but even at smaller lead times, the forecasts can be used to support decision-making to mitigate the impacts of groundwater droughts.

Minor points

Line 25: sentence beginning at end of line - reference style should be Author (Date)

Change made.

line 63: The groundwater level prediction model is referred to as a 'lumped-parameter' model, which seems an odd characterisation. I wonder whether it could be better described as a conceptual model for groundwater level prediction, as the idea of 'lumped-parameters' suggest to me that a common set of model parameters is used for many locations.

Thank you for this comment. We used this wording in previous work (Collenteur et al. 2023) and prefer to stick to this. The wording “conceptual model” is confusing for groundwater hydrologists and hydrogeologists, as this refers to a specific phase in the modeling process of a numerical model. We refer to the introduction of Collenteur et al. (2023) where this terminology is explained further.

Collenteur, R. A., Moeck, C., Schirmer, M., and Birk, S. (2023). Analysis of nationwide groundwater monitoring networks using lumped-parameter models. Journal of Hydrology, 626:130120.

Figure 6: I wonder if there could be a better color scale that would highlight subtleties in forecast skill patterns.

We will test a few different color scale to see if we can improve this.