Reviewer 3

The authors present a new approach to forecast groundwater levels in observation wells, based on measurement in the observation wells prior to the forecast and on forecasts of meteo data. They test their approach for up to 32 days into the future for 6 wells in Switzerland with different characteristics. They do a thorough job in assessing the skill of their forecast using historical data and clearly demonstrate that their forecast approach outperforms naive forecasting approaches. The authors also estimate and assess the uncertainty of their forecast. The approach works well for most of the wells considered. I really like the paper and am not aware of a more thorough study on forecasting groundwater levels. But I think the paper can be improved in some places by better explaining the chosen approach. Some of the language and the conclusions can probably be sharpened a bit. I have quite a few small suggestions, but none of them are really major.

We thank the reviewer for his/her thoughtful comments and many suggestions for improvements to our manuscript. We will use these to further clarify the manuscript, particularly the chosen methodology and better explain and discuss the results.

Minor comments

1. Is there any reason why the lead time is chosen to be 32 days? It seems a bit of an odd choice (but maybe common in the forecasting community). Why not 28 days (4 weeks), 30 days (for those of us liking the decimal system). But 32? It is 4 weeks and 4 days. And it is a power of 2. But other than that?

This is indeed a somewhat odd number, but common in the forecasting community because this was the forecast range for short-term forecasts from the ECMWF. We will add the origin of this number to the manuscript.

2. The first verification of the forecast is the Spearman coefficient. Why Spearman and not Pearson, which I would find a bit more logical and convincing choice. On line 204 it is not indicated what is a good number. On line 10 in the abstract the authors report 0.77, so apparently they think that is a good number. Please explain (maybe with a few references).

Both Spearman and Pearson are commonly used in forecasting studies. Pearson assumes a linear relationship, which we did not want to assume. We will compute the Pearson correlation and provide this information in the appendices.

3. Lines 11-12. Make explicit in the Abstract what the difference is between “driven by recent meteorology and climatology” and “meteorological forecast data”, as this is a main conclusion, but not understandable from these sentences (it is understandable when reading the entire paper).

Thank you for this comment; we will update the abstract to clarify the difference.

4. Fig. 1. I find Fig. 1 somewhat confusing. Some ideas: Put calibration first and forecasting below that. Add the word “calibration” above the first box and the word “forecasting” to the second box. The word “orange” in the caption should probably be “green”.

Good suggestions; we will improve this figure for the revision.

5. Line 112-113. “In forecasting mode … in a post-processing step”. Please explain what is done.

This is explained later on in the text, but we agree this is confusing here. We will change the text to make clear this is explained later in the text and link to the section.

6. An AR1 noise model is applied in an attempt to transform the residuals to uncorrelated noise. I assume it reduced the autocorrelation, but probably didn’t eliminate it (especially since daily data was used). Please explain that either the autocorrelation was removed or that the autocorrelation was just reduced, but it was still assumed the alpha parameter could be used in forecasts.

Excellent point. The models are calibrated to weekly data, and removed most of the autocorrelation. Still, for some models there is autocorrelation left, and it is assumed that the alpha parameter could be used. We will further explain this in the text and refer the reader to materials in the supplementary materials with the results from the autocorrelation analysis, also in relation to comments from the other reviewers.

7. Line 136. Why 51? Seems an odd number as well. I guess it is 3 times 17, but other than that? Why not just 50?

This seems indeed a strange number, but it is what is provided by the ECMWF. We will state that this number is dependent on the input data.

8. Line 150. Isn't the "forecast error" the "forecast variance"? Furthermore, on this line "h" is the time step number, while in Eq. 2 "h" is the lead time. Also, Eq. 3 is only for constant time steps while Eq. 2 is for variable time step. Please fix.

This is indeed the variance of the forecast. We will adapt the text in lines 140-155 to be consistent and have the formula work with variable time steps, and correct the formulas.

9. Line 214 and further. Please explain what the "spread-to-error" ratio is, as it is not commonly used in groundwater hydrology. Do you compute sigma for each ensemble and then take the mean? Or something else? Furthermore, shouldn't the standard deviation of the forecast be compared to the standard deviation of the data? That gives an idea how much the method has added.

Good point. We indeed take the mean of the variance and the mean of the error to compute this metric. We will explain this metric in more detail, as it is indeed not used in groundwater hydrology a lot.

10. Lines 289-290. As the majority of the prediction uncertainty comes from the meteorological forecast rather than the parameter uncertainty, is your forecast less skillful when not using the uncertainty in the parameters at all?

This is an interesting comment. We suspect the parameter uncertainty is indeed only a small portion of the total uncertainty. We will compute this portion to investigate this and report this in the manuscript.

11. Figure 3. There are no lines (or measurement points) outside the 95% ensemble forecast. Does that make sense? Or is this not the 95% interval? Please clarify.

There are actually some lines outside the 95% prediction interval for the well in Kestenholz, but this may be difficult to see from the figure. We will improve the figure by adding a more descriptive legend and change in color scheme, and adapt the figure caption to better explain what is shown.

12. Line 315. "As errors in the meteo input accumulate" and again the word "error" on line 318 and line 325. Why are there errors in the meteo data? You mean variations?

The errors in the forecasted meteorological data is meant here, but this was not clear. We will rewrite this sentence to improve the clarity.

13. Lines 394-400. I find this less clear than the conclusion in the previous paragraph. The first sentence is based on what? Figure 6? Really only Lamone, Gossau and Trub in Fig. 5. And then in the last sentence "in winter" so now suddenly back to winter? Please clarify this discussion.

This is indeed based on the results shown in figure 6. We will update the text to be clear and precise.

Editorial comments

1. On line 6 the authors talk about 32-day lead time, and on line 10 about one-month ahead. Stick with the 32 days as that is investigated.

Change made.

2. Remove the last sentence of the Abstract. Doesn't add anything.

Change made.

3. Line 25. "Sherrer et al." should not be inside the parentheses.

Change made.

4. Line 29. "discharge" -> "recharge"?

No change made, the referenced study really shows a potential decline in discharge from groundwater (as in exfiltration).

5. Line 70. Remove. Doesn't add anything.

Change made.

6. Line 92. "model" -> "Pastas model". Line 94 "applied model" -> "Pastas model". Line 97. "model" -> "Pastas model"

Change made.

7. Line 99-100. "when subtracting the measurements from the simulation the residuals are obtained". Isn't that the other way around?

Good catch, this is indeed the case. We changed the text accordingly.

8. Line 125. "These standard errors". I assume the entire covariance matrix was used?

Correct, no change made.

9. Line 158. Please give reference for "law of total variance".

A reference will be added.

10. Line 163-168. "The performance … approaches". This whole discussion should be moved to the "Discussion" section.

We will move this part to the discussion and slightly rewrite it to fit there.

11. Line 229. Space missing at end of sentence.

Change made.

12. Line 236. Table 1 is not in the appendix, but on the next page. Again on line 253.

Change made.

13. Table 1. DTW, isn't that variable? Is this the mean? Also, please provide units for all quantities in the caption. And finally, what is a snow day? A day that snow falls or a day that there is snow on the ground (which may be more relevant for groundwater)?

DTW is indeed the average. We will update the caption to include the units. A snow day is defined here as every day with precipitation where the temperature is below zero. We will clarify this in the text.

14. Line 254. The “yearly” precipitation. (please add “yearly”). Then on line 255 the word “between” appears twice.

We removed “between” and opted for “annual” instead of yearly.

15. Table 2. Confusing that halfway there is a shift from MAE to NSE. Can the table be improved?

We will try and improve the table to better distinguish between MAE and NSE.

16. Line 307. “perfect git” -> “perfect forecast”.

We replaced “perfect fit” with “perfect forecast”

17. Fig 4. Fix overlapping numbers on horizontal axis. In caption: The orange line “indicated” -> “indicates”.

We updated the figure and the caption as suggested.

18. I think something went wrong with this figure as the label on the vertical axis appears in strange places. Also in caption “increasing” -> “larger” and “declining forecast quality” -> “worse forecast”.

We updated the caption and fixed the figure.

19 Line 351-352. “combination … forecasted” -> “EGP”.

Change made.

20. Fig. 6. It is unclear what is compared to what in which column, and I think the middle column is never even mentioned.

We will improve the description of the figure both in the text and the caption to address this comment.

21. Line 385-386. “for small lead times” but then at end of sentence “for lead times up to 32 days”. Which one is it? Also 2 lines down “did perform” -> “performed”

This is indeed unclear. We will rewrite this sentence. It is small lead times for fast responding systems and up to 32-days for slow groundwater systems.

22. Line 416. “likely explains” -> “likely contributes to”?

Change made.

23. Line 449. “forecasts with high quality”. That is pretty vague and probably too much. Howbout “skillful forecasts”, which is more precise and also used in the Abstract.

Thank you for this comment, we will use skillful as that is indeed more accurate.