Saraceni, M., Gazzetta, G., Silvestri, L., Brunone, B., Meniconi, S., and Bongioannini Cerlini, P. (2025). Seasonal forecasting of water table elevation in shallow unconfined aquifers with case studies in the Umbria Region, Italy. Submitted to Hydrology and Earth System Sciences (HESS), EGUsphere (ID 234233, egusphere-2025-3677)

We thank the Editor and the Reviewers for their careful reading of our manuscript and constructive comments. In the revised version, we tried to address all the issues raised. As reported below, in most cases we have followed the Reviewers' suggestions, while in a few we have provided our reasons for leaving the manuscript unchanged. Below, a detailed, point-by-point response is provided. It is worth noting that we corrected some typos and made some changes to Fig. 1.

Reviewer #1

We thank Reviewer #1 for the attention she/he paid to our manuscript. In the revised version we tried to address all the issues raised. As reported below, in most cases we have followed Reviewer #1's suggestions, while in a few others we have provided our reasons for leaving the manuscript unchanged. It is worth noting that we corrected some typos in Fig. 1

The manuscript is fairly well written, and tries to tackle a relevant research question – establishing simple models for groundwater level predictions allowing to perform seasonal forecasts of this important hydrological variable.

We thank Reviewer #1 for pointing out the relevance of the topic of the paper.

However, I do have major concerns about the validity of the reported work in the current form. I hope these can be addressed by the authors, to ensure the relevance of their work.

As anticipated, we have committed ourselves to improving our paper in this review.

In the conventional understanding, "numerical" groundwater models are physically-based, i.e. try to encode physical processes e.g. via solving governing equations of Darcy's flow or similar. Your model should be considered an empirical or statistical model, similar to the widely used nonlinear transfer function noise models - and it should be discussed in this context.

We fully agree with Reviewer #1 that the Introduction needs to be substantially restructured. Accordingly, in the revised version, the characteristics of the main approaches available in the literature have been briefly discussed, and the proposed model has then been framed within this context.

Another question is: Why do you not simply use those very well-established TFN models? See them e.g., implemented in the Python package pastas https://ngwa.onlinelibrary.wiley.com/doi/full/10.1111/gwat.12925. The empirical formula used does not allow for memory effects in groundwater levels that go beyond the monthly scale – they assume direct scaling from each month's recharge to the same month's groundwater level – again, this limits the applicability of the model to very shallow aquifers without impact from slower-reacting groundwater flow or recharge. Again, the question arises why you do not use established TFN models, which, in many implementation handle this issue, i.e. include past stresses in their response function.

As discussed in the paper, one of the main reasons for using data from reanalysis is that the seasonal forecasting system associated with the modeling system producing reanalysis also provides forecasted soil moisture. Such a quantity can then be used to forecast the water table elevation using the proposed model. Further key advantages of the proposed model are: i) many local, state, and federal agencies maintain databases of the water table elevation, h_w, which is an easy-to-measure quantity, and ii)

several global reanalysis datasets produced by different agencies (Bongioannini Cerlini et al., 2023), such as ERA5, provide near-global coverage and are freely accessible, although they differ in spatial resolution, temporal coverage, and modeling frameworks. Additionally, in response to the Reviewer #1 comment, we have included mentions to an analysis of the time lag between computed fluxes from the reanalysis and water table elevation based on previous work (Bongioannini Cerlini et al., 2021, 2023). Our findings indicate that the piezometers under study are associated with relatively shallow aquifers, resulting in a maximum time lag (τ_{max}) close to 0 days for almost all piezometers. This suggests that the aquifers exhibit minimal memory effects. For clarity, this information has been explicitly added to the piezometer description section of the manuscript.

Having said that, I struggle seeing that sound conclusions can be reached based on results from a single wells. Groundwater level hydrographs and responses to climatic input (or, related soil moisture and derived recharge) are highly diverse. I strongly suggest to apply this method for various wells, which should be a doable task, potentially spanning different aquifer settings etc., assuming that this will yield interesting insights both for the authors and any reader of the manuscript.

We thank Reviewer #1 for her/his comment because it allowed us to explore the applicability of our methodology to different situations. We fully agree with Reviewer #1 that groundwater level dynamics and their response to climatic forcing can vary substantially across space, depending on local hydrogeological conditions, aquifer properties, and anthropogenic influences. Conclusions based on a single well would indeed be insufficient to claim general applicability of the proposed approach. For this reason, the methodology has been tested on a broader set of piezometers across the Umbria Region concerning different aquifers, all belonging to the regional monitoring network and selected following the same objective criteria (data continuity, absence of strong anthropogenic disturbances, shallow unconfined aquifer conditions). Now, the P1-Pistrino piezometer and the P6-Gubbio piezometer are presented in detail as representative examples, but the same analysis has been applied to all piezometers, reported in the "Supplementary Information" section. We chose P1-Pistrino and P6-Gubbio as representative piezometers because they have different depths, different seasonal variability, and different lengths of data availability. We have extended Figures 3 to 8 for P1-Pistrino to include P6-Gubbio, and we have also extended Table 1 and Table 3 with data for all piezometers. In the "Supplementary Information", we have included Figures and data concerning the performances of procedure OPT#1 for all the additional piezometers not included in the main manuscript. We have therefore modified the entire structure of the results text to include the other piezometers, in particular P6-Gubbio, with regard to OPT #2, to the sensitivity test analysis that Reviewer #2 proposed, and to the "naive" forecast analysis that Reviewer #1 proposed. Accordingly, we slightly changed the title of the paper (present title: "Seasonal forecasting of water table elevation in shallow unconfined aquifers **with case studies** in the Umbria Region, Italy").
The additional piezometers span different hydrogeological settings within Umbria, characterized by varying mean water table depths and amplitudes of seasonal fluctuations. Across this set of wells, the proposed framework shows consistent behavior in terms of seasonal predictability and forecast skill, supporting the robustness of the approach beyond a single location. We agree that extending the analysis to an even wider range of aquifer types and climatic settings would provide further insights. This is a natural next step of the research and will be pursued in future work, building on the encouraging results obtained for the Umbria Region.

You deselect wells that show (i) human influence, ... I guess I understand (i) for reasons of simplicity/model development.

We thank Reviewer #1 for her/his comment because it allowed us to discuss the approach we have followed in more detail. As you pointed out appropriately, the exclusion of observation wells with no or a very limited human influence simplifies the problem. It is worth noting that no withdrawal is executed at the considered observation wells, as they belong to a monitoring network. Consequently, the measured considerable variations in the water table elevation in the observation wells not considered in this paper are the result of withdrawals, whose quantity is not known, as well as the temporal placement – carried out elsewhere at an unspecified distance from the observation well. However, the fact that this assumption simplifies the problem is not the only reason, and we would say it is the least important, as clarified just below (at Reply marked with **).

and (ii) limit yourself to a quite narrow (and shallow) range of groundwater depths (lines 65 ff).

The "quite narrow (and shallow) range of groundwater depths" is the consequence of the assumption that the main recharge mechanism of the aquifer is the flow through the unsaturated zone. Moreover, data that allow us to evaluate the flow toward the aquifer are the soil moisture at a depth of about 3 meters. In our opinion, this assumption and the used data are not equally appropriate for analyzing the response of deep aquifers. A future research objective is the possible extension to deep aquifers.

However, I would postulate that in the end you are in particular interested in forecasts in exactly such disturbed aquifers – you yourself set the scene in the introduction with the context of water management strategies. There is little or no management in undisturbed wells. Also, please note that there are enough relatively simple modelling approaches – such as the mentioned TFN - that also allow easy integration of interference such as pumping – e.g. https://doi.org/10.1016/j.jhydrol.2016.01.042 (Hocking and Kelly, 2016)

Reply (**): The main reason we excluded observation wells with considerable variations in water table elevation due to withdrawals (executed elsewhere) is that the poor knowledge of user behavior (in terms of location, entity, and timing of withdrawals) would have made analyzing aquifer behavior more difficult and less reliable. Conversely, simulating the behavior of the aquifer in areas with no (or negligible) withdrawal makes the proposed procedure (Fig. 1) much more effective. Furthermore, the resulting behavior of the water table elevation is of great interest to aquifer managers. In fact, if the withdrawal regime does not change, we can predict the behavior of the water table elevation over the next six months according to environmental conditions (i.e. soil moisture resulting from rainfall and infiltration given by the forecasting system). Thank you for suggesting to us the paper by Hocking and Kelly, (2016) that can be used to increase the number of observation wells.

Furthermore, limiting yourself to shallow wells, you once more exclude relevant aquifers – which, in many cases, for drinking water or other water supply, originate from deeper aquifers/wells due to water quality concerns etc. Your choices severely limit the applicability of your method within the context you yourself set out – at the very least this needs to be discussed.

We fully agree with Reviewer #1 about the great relevance of deeper aquifers. However, we have clearly indicated the limits of our procedure – deriving from the above-mentioned assumptions – for the sake of intellectual honesty. In the revised version, we have motivated our approach in more detail. A future research objective is the possible extension to deep aquifers.
For the sake of clarity we report below the text in the revised version concerning this issue:
In the revised Introduction:
"The limits of applicability of the proposed procedure for forecasting the water table elevation, particularly in terms of aquifer depth and characteristics of the piezometers to be used, derive from the assumptions made, which are discussed in detail in Subsection 3.1."
In the new Subsection 3.1:
"Subsection 3.1: Selection of piezometers and reanalysis datasets and evaluation of flux towards the aquifer: In the region of interest, within a procedure for the identification of available high-quality environmental data (Silvestri et al., 2022), the selection of the piezometers can be based on the criteria discussed in (Bongioannini Cerlini et al., 2021). In particular, the presence of disturbances induced by human activities, and the mean depth of the water table, D, relative to ground level, must be considered.
... Regarding the second criterion, the mean water table depth must range between 4 m and 10 m. The lower bound is related to the length of the soil column, d, explored in the hydrological model providing the reanalysis. In this perspective, it has to be pointed out that ERA5 is characterized by the largest value of d, equal to 2.89 m. The upper bound represents a reasonable maximum depth for shallow aquifers, where infiltration behaves as the primary recharge mechanism. According to Seibert et al. (2003), this depth range ensures that the interaction between the vadose zone and the aquifer is largely unidirectional."

Finally, why do you not use precipitation (or net precipitation, i.e., P–ET) as input directly? Or at least test this as another alternative. After all, soil moisture in reanalysis datasets is a modelled product itself, with inherent uncertainties,

based on the climate data of the reanalysis. Especially as global-scale reanalysis must contain soil parameterization with very limited validity at point scale.

We thank Reviewer #1 for this insightful comment and fully agree that precipitation (or net precipitation, P–ET) represents the primary climatic driver of groundwater recharge. The choice of using soil moisture–derived fluxes instead of precipitation as a direct input was deliberate and motivated by both physical and practical considerations.

First consideration: soil moisture acts as an integrative state variable that naturally filters precipitation through land-surface processes such as infiltration, evapotranspiration, runoff generation, and soil water storage. In shallow unconfined aquifers, groundwater recharge is not driven by precipitation alone, but by the fraction of precipitation that effectively percolates through the vadose zone. Using soil moisture, therefore allows us to implicitly account for these nonlinear processes and for the temporal memory of the soil–aquifer system, which is particularly relevant at monthly and seasonal time scales.

Second consideration: precipitation forecasts—especially at seasonal lead times—are known to be highly uncertain, particularly in Mediterranean regions (Saraceni et al., 2023) where recharge often depends on short, high-intensity events that are poorly captured by global forecast systems. Soil moisture reanalysis and forecasts, while themselves model-derived, benefit from data assimilation and from the integration of multiple surface and subsurface processes, resulting in a smoother and often more predictable signal at seasonal scales. This aspect has been demonstrated in previous studies and is one of the reasons why seasonal soil moisture forecasts often show higher skill than precipitation forecasts over the same regions and time scales (Silvestri et al., 2025). We tried to include these comments in the revised version of the paper (e.g., in the Discussion Section).

Third consideration: we acknowledge that soil moisture in global reanalysis datasets is affected by uncertainties related to land-surface parameterizations and soil properties, which may not be fully representative at the point scale. However, our approach exploits fluxes derived consistently from the same modeling framework (ERA5–SEAS5) and is calibrated locally using observed groundwater levels. This calibration step effectively absorbs part of the structural and parametric uncertainties associated with the land-surface model.

Finally, we agree that testing precipitation- or P–ET-based formulations would be a valuable complementary analysis. However, this would represent a distinct modeling framework requiring additional assumptions on recharge efficiency, temporal aggregation, and lag structures. To maintain focus, the present study concentrates on demonstrating the feasibility of a soil-moisture-based seasonal groundwater forecasting approach. A systematic comparison between precipitation-based and soil-moisture-based predictors is planned as a natural extension of this work. We have clarified these motivations in the revised manuscript in the Discussion section to better justify the methodological choice and to explicitly acknowledge the limitations and potential alternatives highlighted by Reviewer #1

"Indeed, soil moisture–derived fluxes were used as predictors because soil moisture integrates precipitation through the land-surface processes such as infiltration, evapotranspiration, runoff, and soil water storage. In shallow unconfined aquifers, groundwater recharge depends on the fraction of precipitation that percolates through the vadose zone, so soil moisture naturally accounts for these nonlinear processes and the temporal memory of the soil–aquifer system, particularly at monthly and seasonal scales. As mentioned Seasonal precipitation forecasts are often highly uncertain, whereas soil moisture reanalysis and forecasts benefit from data assimilation and the integration of multiple surface and subsurface processes, resulting in a smoother and more predictable signal. Although uncertainties remain due to land-surface parameterizations, the use of a consistent modeling framework (ERA5–SEAS5) calibrated with observed groundwater levels might help in mitigating these types of errors. Precipitation, or net precipitation (P - ET, Precipitation - Evapotranspiration) approaches could, however, provide complementary insights (Almanaseer et al., 2014; Mackay et al., 2015); thus, even if the present study focuses on demonstrating the feasibility of soil-moisture-based seasonal groundwater forecasts, systematic comparisons are planned for future work."

Some more detailed comments:

Line 38: "Recently, as an alternative option, [. . .]" There are many alternative options. Please elaborate a bit more.

In the Introduction of the revised version of the paper, we have discussed about alternative options. Moreover, we tried to link the proposed model for simulating the behavior of the water table elevation to different models used for evaluating the aquifer

recharge. In particular, "Recently, for simulating the behavior of h_w, an empirical model based on the joint use of water table elevation measurements, from piezometers, and soil moisture data, from reanalysis, has been proposed in Bongioannini Cerlini et al. (2017, 2021). Referring broadly to the aforementioned distinction between *from above* and *from below* models, it can be affirmed that such a model can be framed as a sort of synthesis of the two. In fact, as discussed in more detail later in this paper, soil moisture data and related flux toward the aquifer through the vadose zone, provided by reanalysis as a result of global atmospheric models and in situ observations, can be considered of the *from above* type. Oppositely, water table elevations, measured at piezometers, are *from below* data."

Line 40: "reanalysis covers all the world and is fully open access" There are different reanalysis datasets, and this certainly does not apply universally

We thank Reviewer #1 for this comment and agree that the original statement was overly general. In the revised manuscript, we now clarify that several global reanalysis datasets with near-global coverage and open or freely accessible data policies are currently available, including ERA5 (ECMWF), ERA5-Land, MERRA-2 (NASA), and JRA-55 (JMA). These products differ in spatial resolution, temporal coverage, and land-surface modeling, but they all provide global-scale reanalysis data that are accessible to the scientific community. We have therefore rephrased the sentence to explicitly refer to ERA5 and to acknowledge that not all reanalysis datasets universally share global coverage or open-access characteristics:

"Secondly, several global reanalysis datasets produced by different agencies (Cerlini et al 2023), such as ERA5, provide near-global coverage and are freely accessible, although they differ in spatial resolution, temporal coverage, and modeling frameworks."

Performance evaluation: I am missing a benchmark performance of the model, without the effect of SEAS5 forecast uncertainty. Best would probably be the performance of the model forced with ERA5 input directly.

According to Reviewer #1's suggestion, we added the performance evaluation of ERA5 for each piezometer in Section 5, with a new Table, Table 4, which shows the same metrics as Tables 5 and 6, and the now Table 9 for comparison. A comparison is presented in the new Section 6.1:

"A comparison between the ERA5-based simulations of water table elevation (Table 4) and the SEAS5-driven forecasts (Table 5) highlights the expected reduction in skill when moving from reanalysis-based simulation to prediction mode for both piezometers. For P1-Pistrino, the ERA5-driven simulation yields a high skill (KGE = 0.70 and R = 0.81 in Table 4) associated with relatively low errors (RMSE = 0.34 m in Table 4). At the same time, the seasonal forecasts retain satisfactory performance at short lead times, with KGE values of approximately 0.52 at lead time 1 and 0.50 at lead time 2, and correlation coefficients remaining above 0.70 (Table 5). However, the RMSE increases to about 0.5–0.6 m and the MAE to approximately 0.4–0.5 m, indicating a moderate loss of accuracy. Similar consideration can be made for P6-Gubbio. The ERA5-based simulation shows good agreement with observations (KGE $\simeq$ 0.70 and R $\simeq$ 0.80), with RMSE values of the order of 0.9-1 m, while the seasonal forecasts exhibit slightly higher RMSE and MAE (Table 5), but lower KGE values of about 0.60 at lead time 1 and 0.58 at lead time 2. Despite this degradation, the persistence of positive KGE values, relatively high correlations, and limited error growth at short lead times indicates that a substantial fraction of the groundwater signal captured by ERA5 is retained in the seasonal forecasts. The results of the water table elevation forecast for the other considered piezometers in the Umbria Region are reported in the Supplementary Information."

Furthermore, how skillful is the model really? A reasonable KGE at monthly scale for this kind of time series can probably be achieved very easily, as, in principle, we see the same/a very similar seasonal patterns with winter (spring) highs and summer (autumn) lows repeated every year. Those originate from the seasonal patterns of (net) precipitation, which the seasonal forecast easily reproduces. As a benchmark, the performance of a simple climatology should be considered, or similar. I.e. calculate the average water table across all January, February, etc. values to obtain the climatology, and then compare this against the actual monthly timeseries. Any forecast model has to be better than this "forecast" that

can be obtained "for free".

We agree with Reviewer #1 that, at monthly time scales, groundwater level time series are strongly dominated by a seasonal cycle, mainly driven by the climatological pattern of precipitation and recharge. As a consequence, relatively high values of correlation-based metrics or KGE can indeed be achieved by very simple benchmarks, such as a monthly climatology, and any forecasting model should demonstrate added value with respect to such "free" predictors.

To explicitly address this point, we have now introduced a climatological benchmark, constructed as suggested by the reviewer. For each calendar month, we compute the mean observed water table elevation over the reference period (2001–2016), thus obtaining a monthly climatology. This climatological sequence is then compared against the observed monthly water table time series over the forecast evaluation period (2012–2020 for P1-Pistrino and from 2015 to 2020 for P6-Gubbio), and the same skill metrics used for the forecast model (including KGE, RMSE, MAE, and correlation) are calculated.

Now in the results Section 6.1, there is a whole paragraph about this analysis:

"To explicitly assess the added value of the seasonal forecasting framework relative to the simplest baseline predictor, we constructed a monthly climatology benchmark for the two representative piezometers by averaging observed water table levels for each calendar month over the calibration period (2001–2016). This climatology was then used as a deterministic forecast for the verification period (2012–2020 for P1-Pistrino and 2015-2020 for P6-Gubbio). The results of the climatological analysis are reported in Table 6. The "climatological" forecasts lead to slightly higher errors than the SEAS5-driven forecasts at short lead times. For P1-Pistrino, the climatology yields KGE = 0.57 and correlation $\simeq 0.72$, and higher RMSE than the model forecasts at lead time 1, but comparable at lead time 2. For P6-Gubbio, similar considerations can be made, with RMSE and MAE from the climatology that are higher than the seasonal forecasted values (comparison with Table 5). This indicates that the proposed framework predicts departures from the mean seasonal pattern rather than merely repeating it, at least at short lead times."

It is important to note that the anomalies in seasonal forecasts, compared to the climatological mean observation, may be affected by inherent biases (see also computation of CRPSS in the Supplementary Information), potentially linked to the different (coarser) spatial resolutions of seasonal forecasts compared to both the ERA5 reanalysis and local observations. This mismatch in resolution could lead to systematic biases in the representation of groundwater anomalies, especially in areas with complex hydrogeological settings. To address this issue in future work, we suggest exploring a two-step bias adjustment approach: first, adjusting the meteorological inputs (as already implemented in this study), and second, applying an additional bias adjustment directly to the simulated groundwater levels after their calculation. This second step could help correct residual biases in the groundwater level forecasts, improving the alignment between model outputs and observations, and ultimately enhancing the reliability of the forecasting system. We underlined this in the discussion Section of the main manuscript.

Table 4, 5, 6: Performance values should probably be provided with to valid digits

This has been corrected as Reviewer #1 suggested.

Reviewer #2

In this study a system is developed to provide seasonal outlooks of groundwater levels up to 7 months into the future. The method uses soil moisture estimates from the SEAS5/ERA5 systems to predict the groundwater levels with natural fluctuations. The method is tested on a single monitoring well in Italy using monthly mean groundwater levels. The results show that reasonably accurate forecasts can be provided, but large deviations for higher groundwater levels in springtime. The use of soil moisture rather than meteorological input directly is an interesting approach to forecast groundwater levels, which I did not encounter before. Below I outline some major concerns, with minor line comments further down.

Major comments:

The introduction section does mention any review of studies attempting to generate seasonal groundwater levels outlooks/forecasts, while literature is available. I recommend including a one or two-paragraph review of existing work and clearly stating what the proposed methodology might add to that work (i.e., using different input data – soil moisture).

According to Reviewer #2's request, in the Introduction of the revised version, the most recent contributions in literature about the forecast of water table elevation are briefly illustrated. Moreover, in the revised version, the most important contribution of the methodology proposed in the paper is pointed out: the fundamental role of global atmospheric models in predicting groundwater levels, as both reanalyses and forecasting systems are linked to them. Below is the relevant text from the revised version:

"To highlight the relevance of the proposed approach, it is worthy of pointing out that hydrological prediction literature shows the relative "infancy" of groundwater level forecasting with respect to the one of streamflow and surface water level (Mackay et al., 2015; Collenteur et al., 2025). As an example of available approaches, it is worth mentioning the one proposed in Mackay et al. (2015), where the probabilistic forecasting of h_w up to five months approach is based on the state-of-the-art GloSea5 multimember seasonal forecasts of rainfall produced by the UK Met Office and lumped conceptual groundwater model Aquimod (Mackay et al., 2014). In Huang et al. (2020), for assessing groundwater level, a procedure based on weather forecasts as inputs of WASH123D hydrological model (Yeh et al. 2011) is used. In Robertson et al. (2024), for a case-study catchment, the hydrological model is first refined to improve streamflow and groundwater level predictions. Successively, errors and uncertainty in groundwater level predictions are reduced and quantified according to the FoGSS approach proposed by Bennett et al. (2016). As pointed out in the successive Sections, the novelty of the procedure proposed in this paper gives global atmospheric models a fundamental role in predicting groundwater levels, as both reanalysis and forecasting systems are linked to them."

The methodology is tested on a single well, which makes it difficult to generalize the results, which also shows clear seasonal fluctuations and probably a good predictability. It would be good if the Authors could extend the work to include a few more wells, to see if/how the method generalizes.

We thank Reviewer #2 for her/his comment because it allowed us to explore the applicability of our methodology to different wells in different conditions. Conclusions based on a single well would indeed be insufficient to claim general applicability of the proposed approach. For this reason, the methodology has been tested on a broader set of piezometers across the Umbria Region, all belonging to the regional monitoring network and selected following the same objective criteria (data continuity, absence of strong anthropogenic disturbances, shallow unconfined aquifer conditions). Now, the P1-Pistrino piezometer and the P6-Gubbio piezometer are presented in detail as a representative example, but the same analysis has been applied to all piezometers, reported in the "Supplementary Information" section. We chose P1-Pistrino and P6-Gubbio as representative piezometers because they have different depths, different seasonal variability, and different lengths of data availability. We have extended Figures 3 to 8 for P1-Pistrino to include P6-Gubbio, and we have also extended Table 1 and Table 3 with data for all piezometers. In the "Supplementary Information", we have included figures and Table for all the performances of procedure

1 (OPT #1) for all the additional piezometers not represented in the main manuscript. We have then modified the rest of the structure of the results text to include the other piezometer, in particular P6-Gubbio, with regard to OPT #2, to the sensitivity test analysis, and to the "naive" forecast analysis that Reviewer #2 proposed. Accordingly, we slightly changed the title of the paper (present title: "Seasonal forecasting of water table elevation in shallow unconfined aquifers **with case studies** in the Umbria Region, Italy").

The additional piezometers span different hydrogeological settings within the Umbria region, characterized by different mean water table depths and amplitude of seasonal fluctuations. Across this set of wells, the proposed framework shows consistent behavior in terms of seasonal predictability and forecast skill, supporting the robustness of the approach beyond a single location. We agree that extending the analysis to an even wider range of aquifer types and climatic settings would provide further insights. This is the next step of the research and will be carriedd out in future work, building on the encouraging results obtained for the Umbria Region.

The approach to how the groundwater level model is developed is only briefly described (section 2.2), and I found it challenging to follow without any details about what was done here. Things become a bit clearer when the case study is described in Section 3 / Formulas 9 and 10. I would prefer Section 2.2 to be more in-depth, and generalizable to other monitoring wells with different model structures.

According to Reviewer #2's request, in the revised version of the paper, the description of the model used for the simulation of the water table elevation is provided in advance in Section 2. This is believed to make the proposed procedure easier to understand. Relationships used in the Umbria region are given in Section 4 where the case studies are illustrated.

There is no estimation or even discussion of the model parameter uncertainty, which probably is substantial. I think this should at least be discussed, but preferably added to the simulation results.

We thank Reviewer # 2 for this important comment. Following this suggestion, we have now explicitly addressed model parameter uncertainty and incorporated it into the simulation results in a separate Section, Section 5.2, "Parameter sensitivity analysis". In the revised manuscript, the parameter uncertainty was explored through a systematic perturbation of the calibrated model parameters k_{lin}^m and k_{log}^m. Each parameter was independently varied by $\pm 20\%$ and $\pm 10\%$ around their reference values, while keeping the other parameter fixed. These perturbed parameter sets were then used to re-run the water table forecasts of OPT #1. This sensitivity-based analysis allows us to evaluate the impact of plausible parameter uncertainty on forecasted water table levels without introducing additional assumptions on parameter distributions. We have added a dedicated discussion of these results in the revised manuscript in the results section, with the values for all the parameters' changes in Table 7 and the performance assessment for lead time 1 with the new parameters in Table 8 to be compared with Table 4 at the same Lead time. Furthermore, we included the corresponding results for P1-Pistrino and P6-Gubbio in Supplementary Figures S9 to S12. The results show that parameter uncertainty affects both the amplitude and timing of groundwater level fluctuations, particularly at short lead times, but does not alter the overall seasonal behavior or the main conclusions regarding forecast skill, thus making the results for OPT #1 more robust.

The most important part that I am missing is a comparison with other 'naïve' forecasts, such as persistence or climatology. This is a very common practice in the forecasting community and helps identify the source of the forecast skill. I think adding such forecasts from simpler systems is essential to truly understand if the newly developed system brings additional forecast skill. I recommend adding such naïve forecasts and adding skill scores (i.e., CRPSS) to investigate this. I suspect simpler systems without the soil moisture input data, particularly when taking annual seasonality into account, might perform similarly.

We thank Reviewer #2 for this important comment and fully agree that the use of naïve benchmarks is essential to correctly interpret the source of forecast skill. In the revised manuscript, we explicitly introduced a monthly climatology benchmark and compared its performance against the proposed forecasting system (Table 6). This climatological forecast represents a deterministic "free" predictor that exploits only the mean seasonal cycle.

Eventually, what the proposed framework provides is added value beyond simple naïve forecasts within the first to third lead time as concerns both P1-Pistrino and P6-Gubbio, the representative piezometers.

Following the Reviewer #2's suggestion, we also evaluated probabilistic forecast skill using CRPSS relative to a monthly climatological benchmark. Results (Table S3) show negative CRPSS values, indicating that the SEAS5 ensemble is under-dispersive compared to the climatological distribution. This is consistent with the reduced ensemble spread observed in the forecasts and does not contradict the positive anomaly correlation and deterministic skill shown in the main analysis. However, the CRPSS results help clarify that forecast limitations are primarily related to ensemble dispersion rather than to the absence of predictive information. Indeed it is important to note that the anomalies in seasonal forecasts, compared to the climatological mean observation, may be affected by inherent biases as confirmed by the negative values obtained by the CRPSS, potentially linked to the different (coarser) spatial resolutions of seasonal forecasts compared to both the ERA5 reanalysis and local observations. This mismatch in resolution could lead to systematic biases in the representation of groundwater anomalies, especially in areas with complex hydrogeological settings. To address this issue in future work, the authors will explore a two-step bias adjustment approach: first, adjusting the meteorological inputs (as already implemented in this study), and second, applying an additional bias adjustment directly to the simulated groundwater levels after their calculation. This second step could help correct residual biases in the groundwater level forecasts, improving the alignment between model outputs and observations, and ultimately enhancing the reliability of the forecasting system. We underlined this in the discussion Section of the main manuscript. Furthermore, in the revised manuscript, we added the CRPSS analysis in the supplementary information and a paragraph to the Results Section 5.1:

"In addition to deterministic metrics, probabilistic forecast skill was evaluated using the Continuous Ranked Probability Skill Score (CRPSS), computed relative to a monthly climatological benchmark (the definition of CRPSS is given in the supplementary Information and results are reported in Table S3 of the supplementary information). For both P1–Pistrino and P6–Gubbio, CRPSS values are close to 0 but negative across lead times, indicating that the SEAS5 ensemble spread is narrower than the climatological distribution. This behavior reflects the limited ensemble dispersion. Thus, the forecasts retain skill in reproducing the timing and sign of groundwater level anomalies, while underrepresenting their full amplitude."

Minor comments:

L14: "due to groundwater"?

Data about the relevance of groundwater are from "Protecting groundwater for health: Managing the quality of drinking-water" by IWA Publishing for World Health Organization (2006) (WHO, 2006) and a paper by Aeschbach-Hertig and Gleeson (2012).

L19-20: Please specify or rephrase; the potential is determined by many more factors in terms of quality and quantity than just the water table.

In the revised version of the paper, we modified such a sentence as "From the quantitative point of view, potential of unconfined aquifers – on which attention is focused in this paper – is determined by the elevation of the water table, h_w. However, the relevance of h_w is not negligible also from a qualitative point of view either. In fact, the decrease in h_w due to pumping draws water from more distant and potentially less controlled areas".

L20: Perhaps remove words such as "very" here, and similar instance hereafter.

In the revised version of the paper, we deleted the word "very" in this instance and in the rest of the manuscript.

L31: remove "numerical", there are other model types that can do this.

In the revised version, "numerical" has been removed. As a partial excuse for the oldest writer, the recollection that decades ago, the so-called Hele-Shaw models were used for simulating groundwater flow.

L33: remove "much more challenging indeed"

In the revised version, such a section has been deeply modified.

Figure 1: What is OPT1? Please clarify this in the figure caption.

The description has been added to the Figure's caption according to Reviewer #2's suggestion.

L56: What is done in the other cases?

The authors used the data and variables for the simulation model directly from the ERA5 grid point closest to the piezometer. This is the methodology most commonly used for studies of this type (Bongioannini Cerlini et al., 2021, 2023). For other types of applications, the average of the reanalysis and forecast grid points around the observed point is taken, e.g., for Lake water mass balance modeling (Saraceni et al., 2024). In this case, taking only one grid point uniquely marks the parameters that are used in the model. For clarity, the phrase "in most cases" has been removed from the new text.

L68-69: This is a limitation of the approach that could be discussed. Also, please specify how it was determined that no other (substantial) influences are present?

In the revised version of the paper, such an issue has been addressed in more detail. Precisely, the fluctuations of the water table are analyzed to detect possible significant oscillations due to, for example, irrigation. In particular, as shown in (Bongioannini Cerlini et al., 2021), the existence of an orderly pumping for drinkable water supply or irrigation purposes has been verified not only by consulting local water companies but also by pointing out large water table oscillations occurring in given time intervals (e.g., those when the irrigation is active). The selection procedure also includes the comparison between the behavior of the piezometer under analysis with the one in a neighboring piezometer certified as not affected by pumping. The main reason of excluding piezometers with considerable variations in water table elevation due to withdrawals (executed elsewhere) is that the poor knowledge of user behavior (in terms of location, entity, and timing of withdrawals) would make analyzing aquifer behavior more difficult and less reliable. Conversely, simulating the behavior of the aquifer in areas with no (or negligible) withdrawals makes the proposed procedure (Fig. 2) much more effective. Furthermore, the resulting behavior of the water table elevation is of great interest to aquifer managers. In fact, if the withdrawal regime does not change, the behavior of the water table elevation over the next six months can be predicted according to environmental conditions (i.e. soil moisture levels resulting from rainfall and infiltration given by the forecasting system).

L124: What are the reference values? Please clarify

Both in Section 2, introduced specifically in the revised version, and in Section 4, the meaning of the reference quantities h_w^* and F_g^* has been further specified. In particular, in Section 4, it is shown that h_w^{max} is the 99th percentile of h_w and F_g^{max} is the absolute maximum value of F_g.

L144: instead of "improve calibration", perhaps "improving verification" is meant here?

In the revised version, the authors changed this as suggested by Reviewer #2

L179: No other piezometers fulfill these criteria?

See previous reply on the extension of the analysis to multiple piezometers in the Umbria Region.

Table 4: Please add numbers with 2 decimals

This has been corrected as Reviewer #2 suggested.

L250: I think the raw value says little, and 0.5 meters seems quite a large MAE to me. Perhaps add the variation of the GWL fluctuations for comparison?

We thank Reviewer #2 for this comment. We agree that the interpretation of absolute error values, such as MAE, benefits from being placed in the context of the natural variability of groundwater level fluctuations. For this reason, the standard deviation of observed groundwater level variations for each piezometer is reported in Table 1 (column σ_{h_w}), which provides a direct measure of the typical amplitude of groundwater fluctuations at each site. For the analyzed piezometers, the reported MAE values (approximately 0.5 m for P1-Pistrino and 0.9 m for P6-Gubbio in Table 5 and Table 6) are smaller than the corresponding observed variability, which ranges from about 0.6 m for P1-Pistrino to more than 2 m for P6-Gubbio. The forecast errors are moderate relative to the intrinsic variability of the groundwater system and are therefore consistent with a skillful prediction at seasonal time scales. We have added explicit reference to the groundwater level variability reported in Table 1 when discussing MAE values, thereby facilitating a more meaningful interpretation of forecast errors.

Figure 4: Please consider providing Excel tables with the raw values of all simulations and measurements to reproduce these figures and the table.

This has been provided as Reviewer #2 suggested for P1-Pistrino and P6-Gubbio for OPT #1 procedure on a Zenodo repository with DOI 10.5281/zenodo.18662028. Consequently, a line has been added to the Data availability statement.

L286-287: This statement is a bit vague, please clarify what is exactly meant here.

We agree with Reviewer #2 that this statement is a bit vague in the way it is written. By "more tightly clustered together," we specifically referred to a reduction in ensemble spread (i.e., lower inter-member variability) of the water table level forecasts produced by OPT2 compared to OPT1, particularly at lead times between 2 and 4 months, as shown in Fig. 7b–d. This, however, was a qualitative analysis. In the new version of the manuscript, a direct comparison between the ensemble spread (standard deviation of the ensemble distribution) of OPT #1 and OPT #2 for the two selected piezometers is presented, along with results on differences in ensemble dispersion and uncertainty. Thus, in the revised manuscript, a new Table is presented, Table 8, with results regarding the ensemble spread, and the whole paragraph has been replaced with the new one reported below:

"This is reflected in the ensemble spread analysis in Table 8, which further highlights differences between the two forecasting options. For P1–Pistrino, OPT 2 exhibits a gradual increase in ensemble spread with lead time, from approximately 0.14 m at lead time 1 to about 0.26 m at lead time 6. Compared to OPT 1, this represents a moderate increase in ensemble dispersion at longer lead times, suggesting a more realistic representation of forecast uncertainty. Despite this increase, the ensemble spread remains small, indicating that the forecasts are still under-dispersive. This behavior is consistent with the dynamic recalibration strategy of OPT 2, which reduces MAE and RMSE while allowing uncertainty to grow with lead time. For P6-Gubbio, the ensemble spread under OPT 2 increases markedly with lead time, from approximately 0.30 m at lead time 1 to about 1.04 m at lead time 6. Compared to P1-Pistrino, this reflects the deeper water table and the larger intrinsic variability of this site, although, also in this case the ensemble remains moderately under-dispersive relative to the climatological variability. Overall, the dynamically updated method, OPT2 provides less uncertainty at longer lead times, partially mitigating the limitations observed in OP T1 and generally higher dispersion in the forecasts. This behavior reflects the periodic recalibration of the transfer function, which reduces parameter-related uncertainty and leads to a more coherent ensemble response to the same seasonal

soil moisture forcing, rather than an artificial reduction of forecast uncertainty."

L330: Add: “for this data and case study area” or something similar, because we do not know how well the results generalize based on one well.

This has been corrected as Reviewer #2 suggested.

References

Aeschbach-Hertig, W. and Gleeson, T.: Regional strategies for the accelerating global problem of groundwater depletion, Nature Geoscience, 5, 853–861, 2012.

Almanaseer, N., Sankarasubramanian, A., and Bales, J.: Improving groundwater predictions utilizing seasonal precipitation forecasts from general circulation models forced with sea surface temperature forecasts, Journal of Hydrologic Engineering, 19, 87–98, 2014.

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resources Research, 52, 8238–8259, 2016.

Bongioannini Cerlini, P., Meniconi, S., and Brunone, B.: Groundwater supply and climate change management by means of global atmospheric datasets. Preliminary results, Procedia Engineering, 186, 420–427, 2017.

Bongioannini Cerlini, P., Silvestri, L., Meniconi, S., and Brunone, B.: Simulation of the water table elevation in shallow unconfined aquifers by means of the ERA5 soil moisture dataset. The Umbria region case study, Earth Interactions, 25, 15–32, 2021.

Bongioannini Cerlini, P., Silvestri, L., Meniconi, S., and Brunone, B.: Performance of three reanalyses in simulating the water table elevation in different shallow unconfined aquifers in Central Italy, Meteorological Applications, 30, e2118, 2023.

Mackay, J., Jackson, C., Brookshaw, A., Scaife, A., Cook, J., and Ward, R.: Seasonal forecasting of groundwater levels in principal aquifers of the United Kingdom, Journal of Hydrology, 530, 815–828, 2015.

Saraceni, M., Silvestri, L., Bechtold, P., and Bongioannini Cerlini, P.: Mediterranean tropical-like cyclone forecasts and analysis using the ECMWF ensemble forecasting system with physical parameterization perturbations, Atmospheric Chemistry and Physics, 23, 13 883–13 909, 2023.

Saraceni, M., Brunone, B., Silvestri, L., Meniconi, S., and Cerlini, P. B.: A water mass balance-based procedure using ERA5 Land Reanalysis and level observation to reconstruct the past level changes of closed lakes toward future management, Journal of Hydrometeorology, 2024.

Seibert, J., Rodhe, A., and Bishop, K.: Simulating interactions between saturated and unsaturated storage in a conceptual runoff model, Hydrological Processes, 17, 379–390, 2003.

Silvestri, L., Saraceni, M., and Bongioannini Cerlini, P.: Quality management system and design of an integrated mesoscale meteorological network in Central Italy, Metereological Applications, 29, e2060, 2022.

Silvestri, L., Saraceni, M., Brunone, B., Meniconi, S., Passadore, G., and Bongioannini Cerlini, P.: Assessment of seasonal soil moisture forecasts over the Central Mediterranean, Hydrology and Earth System Sciences, 29, 925–946, 2025.

WHO: Protecting groundwater for health: Managing the quality of drinking-water sources, IWA Publishing for World Health Organization, 2006.