Author's Response

We sincerely thank the editor and the reviewers for their thorough and constructive evaluation of our manuscript. We greatly appreciate the time and effort invested in reviewing our work. The comments have helped us to clarify several methodological aspects and improve the overall presentation of the manuscript.

Below, we respond to each comment in detail. Reviewer comments are shown in black, followed by our responses in blue.

Community Comment

I would suggest to include the purpose of the models in the title, e.g. by adding "for groundwater levels in Germany".

Thank you for the suggestion to further specify the application domain in the title. However, we deliberately chose to keep the title general, as the primary contribution of this study is methodological. The proposed strategies for incorporating static features and for performance evaluation are not specific to groundwater level prediction in Germany, but are intended to be transferable to other regions and to related environmental prediction tasks, such as surface water runoff prediction (see last paragraph of our Conclusions). The German groundwater dataset serves as a comprehensive and well-documented example to demonstrate these strategies, rather than as a limiting case study. To avoid implying restricted applicability, we therefore decided to retain a general title. **No changes made.**

Reviewer #1

The authors compare four different strategies for incorporating static features into global deep learning models. They show that the repetition method generally achieves the best performance but they are less computationally efficient. They also state that the selection of static features is more important than the choice of integration strategies. The manuscript is generally well-presented.

We thank the reviewer for the positive assessment of the manuscript and for the constructive comments, which we believe help to further improve the quality and clarity of the paper. In the following, we address each point raised by the reviewer in detail.

However, I still have the following concerns.

In the Table 1 of Heudorfer et al. (2023), it shows that the time series features were derived from past groundwater level time series until 2011. The training, validation, and test periods for this study are 1991-2007, 2008-2012, 2013-2022, respectively. If the authors directly adopt the time series features from Heudorfer et al. (2023), there might be data leaking issue during validation.

We thank the reviewer for pointing this out. Although the feature definitions follow Heudorfer et al. (2023), all time-series-based static features were recomputed in this study using only information available up to the respective training period. No data from the validation or test periods was used for feature derivation, thus preventing any information leakage**. We have clarified this explicitly in Section 2.4 (Lines 123–125) and added a reference to Appendix 1, where a complete list and description of the time series features is provided.**

Also, while the authors have provided references for the time series features, it is nice to list the time series features in the main text or appendix for readability.

Thank you for that constructive suggestion! **For improved readability, we have added a complete list of the time-series-based static features to the Appendix A (Table A1).**

For the conditional model, it is unclear to me how the output of this layer is split and used to initialize the hidden and cell states of the first LSTM layer in the dynamic branch (Line 167-168). Does it mean that the output is directly used as the initial condition of the hidden and cell states? Please clarify.

We thank the reviewer for pointing out that this description was not sufficiently precise. In the conditional strategy, the static branch output is linearly projected to a vector of size $2H$ (with H being the number of LSTM units), which is then split into two vectors of size H and directly used to initialize the hidden and cell states of the first LSTM layer. **We clarified the conditional architecture in Section 3.3.1 (Lines 173–179), where we now explicitly describe the projection to 2H, the split into h_0 and c_0, and the addition of a second LSTM layer.**

In Line 176, the authors mentioned that there are 10 model initializations. Does it mean that the authors train 10 global models with different initial conditions?

We thank the reviewer for this comment. **We have clarified that the ensemble consists of 10 independently trained global models with identical architecture, data splits, and hyperparameters, differing only in their random weight initialization. Performance metrics are based on the median predictions across these runs (Section 3.3.1, Lines 183–185).**

For the results, the authors that the repetition model performs the best, but they show the results for the 256-neuron model for the repetition model and the 128-neuron model for the other models. It is hard to identify whether the better performance is due to the integration strategy or the more hidden neurons.

We thank the reviewer for this important comment. As described in the manuscript, all integration strategies were implemented using a baseline LSTM size of 128 units to ensure consistency across architectures. For the repetition strategy, we additionally evaluated a configuration with 256 units, as the replication of static features at each time step more than doubles the dimensionality of the dynamic input. In this setting, using a larger recurrent layer constitutes a reasonable architectural adjustment rather than a separate optimization step. Since the 256-unit configuration yielded slightly better performance, only these results are presented for the repetition model. **The corresponding clarification has been added to the revised manuscript (Section 3.3.1, Lines 152–156).**

Specific comments:

Explain labels/legends in Figures 2 and 3.

Thank you for pointing this out**. We have revised the captions of Figures 2 and 3 to clearly explain all abbreviations, model acronyms, and evaluation settings shown in the figures.**

Reviewer #2

This paper explores different strategies for global deep learning models that account for basin and hydrogeological "static" properties and characteristics. These strategies aim to enhance the models' generalization capabilities and overall performance. The authors tested several approaches that differ in how static properties are incorporated into the model and conducted these tests using two types of modeling methods ("in-sample"/training on all available wells, and "out-of-sample"/training 90% of the wells with test on the remaining 10%, test period being equal in both approaches). From the Deep Learning point of view, their study builds on this same technical issue that also emerged from other scientific fields. 4 integration strategies were tested for each modeling approach (in-sample and out-of-sample). The simplest integration strategy (repetition) appeared to perform the best, at the

cost of much lower computational efficiency. However, the authors conclude that other strategies, particularly concatenation and conditional initialization of LSTM weights, deserve thorough consideration as they offer a good balance of performance and computational efficiency. The paper addresses an important issue; it summarizes the usefulness and relevance of existing strategies for incorporation of static features in LSTM in the specific case of hydrogeology. It is a very nice study, clearly written and organized. There are a few points that might be addressed or highlighted in the paper before publication in my opinion.

We thank the reviewer for the thorough and insightful review, as well as for the very positive assessment of the manuscript. We particularly appreciate the constructive comments and suggestions, which we believe help to further strengthen the clarity and relevance of the paper. In the following, we address each point raised by the reviewer in detail.

The paper should provide examples of time series (e.g. in the form of a panel with 4 or 5 of them) without requiring the reader to download them. I think it is important to have a straightforward understanding of the context of hydrological modeling by knowing what ground-truth data looks like. Therefore, it is crucial that a few examples be presented directly in the text. For instance, if most time series consist of almost pure periodic annual variations with constant amplitude through time, expectations regarding the model's performance would not necessarily be the same as for more complex variability. After downloading the time series and briefly examining them, significant differences in statistical properties can be observed (more or less weak trends, very short-term variations, strong amplitude of the water year cyclicity...). Do the authors know if, and to what extent, such differences may play any role in the models' performance: are there some behaviors for which the models systematically perform poorly, or very well?

We thank the reviewer for this insightful comment highlighting the importance of providing a direct impression of the groundwater level time series used in this study. **To better illustrate the characteristics of the observed data, we have added a new figure (Figure 2) in Section 2 to provide example groundwater level time series and illustrate the variability and heterogeneity of groundwater dynamics across wells.**

Regarding the potential influence of different time series characteristics on model performance, we agree that this is an important and interesting question. In the present study, however, we did not conduct a systematic stratification of model performance by time series type or statistical properties. Our primary objective was to evaluate integration strategies for global models intended to operate consistently across large and heterogeneous monitoring networks. In such practical settings, it is typically not feasible to apply different modelling strategies depending on the behaviour of individual monitoring sites. Accordingly, while substantial heterogeneity in groundwater dynamics is present across wells, we think that a detailed performance analysis conditioned on specific time series characteristics is beyond the scope of this work and represents a promising direction for future research. **No changes made.**

Although it is beyond the scope of the paper, it seems like a lot of meteorological inputs was used. To what extent are they all "meaningful" for the application? Have previous studies that used this database conducted SHAP analysis or similar methods to determine which features such models learn from most effectively?

We thank the reviewer for this relevant comment. In this study, we directly used the meteorological input variables provided by the GEMS-GER dataset without performing an explicit feature selection or feature attribution analysis. Our primary objective was not to assess the relevance of individual meteorological predictors, but to investigate how different strategies for integrating static features affect model performance and generalization.

We fully agree that analyzing which meteorological or static inputs are most informative—using methods such as SHAP or related feature attribution techniques—is a highly interesting and important topic. However, such analyses are beyond the scope of the present work and would substantially extend its focus. We therefore consider this a promising direction for future research. **No changes made.**

I think one or two lines on the concept of "meaningful" static features as it is used here would be needed. Here, "meaningful" stands for "informative" if I am not mistaken; maybe this term would be more appropriate.

We thank the reviewer for this helpful suggestion. **Following the reviewer's recommendation, we have replaced the term "meaningful static features" with "informative static features" throughout the manuscript to improve clarity and precision.**

It should be said at the beginning and justified why no hyperparameter optimization was conducted: this is a technical context that should be mentioned and explained (even briefly), especially for researchers who intend to use a similar approach.

We thank the reviewer for this important comment. In this study, we deliberately did not perform an extensive hyperparameter optimization, as our primary objective was to compare different integration strategies under consistent and comparable experimental conditions rather than to maximize predictive performance. All models were therefore implemented using a consistent baseline architecture and a common set of hyperparameters, allowing us to isolate the effects of the integration strategies. **We clarified this in Section 3.3.1 (Lines 161–164).**

Would one-hot encoding be very different than the repetition approach? This is the simplest way to attach an identifier to the wells, so in the framework of this study it would be interesting to recall this.

We thank the reviewer for raising this interesting point. One-hot encoding is indeed a simple and commonly used way to incorporate site identifiers as static inputs. Upon reflection, we note that one-hot encoding and the repetition strategy address different aspects of the model input: while one-hot encoding increases the dimensionality of the static feature space by encoding site identity, the repetition strategy operates along the temporal dimension by replicating static features at each time step of the dynamic input.

In this study, our focus was on integrating physically or environmentally informative static features rather than providing an explicit site identifier. Moreover, the set of static attributes used here is relatively rich (more than 40 variables) and may already provide an implicit characterization of individual wells, i.e., a "soft" identifier, while still being interpretable in terms of hydrogeological and environmental properties. We therefore did not include an additional one-hot well identifier as a separate strategy. We agree, however, that an explicit comparison between identifier-based encodings (e.g., one-hot) and feature-based static descriptions would be an interesting direction for future work. **No changes made.**

Line 149 (LSTM layer size): Why is it 128 when no hyperparameter optimization has been performed? Given that the results of the repetition model were presented for an LSTM layer of size 256, wouldn't it be preferable to present all results with an LSTM size of 256? I am not questioning the relevance of the results presented here, but it is important in my opinion that the presentation of the methodology does not raise any unnecessary questions for researchers interested in developing a similar approach.

We thank the reviewer for this comment. As described in the manuscript, all integration strategies were implemented using a baseline LSTM size of 128 units to ensure consistency across architectures. For the repetition strategy, we additionally evaluated a configuration with 256 units, as the replication of static features at each time step more than doubles the dimensionality of the dynamic input. In this context, using a larger recurrent layer represents a reasonable architectural adjustment rather than a separate hyperparameter optimization step. Since the 256-unit configuration yielded slightly better performance, only these results are presented for the repetition model. **We have revised the corresponding paragraph in Section 3.3.1 to more explicitly explain the choice of LSTM sizes and the reasoning behind reporting the 256-unit repetition model (Lines 153-156).**

Line 106: replace "real" with "actual".

Thank you, **we have replaced the term “real” with “actual” (now Line 107 of the revised manuscript).**

Line 177: Section 3.3.2 and 3.3.3 should be merged in a single "In-sample and Out-of-sample" 3.3.2 section.

We thank the reviewer for this suggestion. We deliberately chose to describe the in-sample and out-of-sample settings in separate sections, as they differ in terms of training setup and evaluation objective. We believe that this separation improves the clarity of the experimental design and therefore decided to retain the current structure. We hope that this organization is acceptable to the reviewer. **No changes made.**

About the quality of environmental static features:

Line 220: I am not sure I understand this point well, which also seems to be about the heterogeneity of physical characteristics, as discussed in the previous point. Regarding the representativeness of the database for sampling general hydrogeological characteristics properly, I believe this can be addressed with general hydrogeological knowledge. Additionally, the spatial coverage and number of wells appear sufficient for consistent sampling of hydrogeological properties.

We thank the reviewer for this thoughtful comment and the opportunity to clarify this point. We agree that two related, but distinct aspects are discussed in the manuscript, and we appreciate the reviewer’s careful reading.

The first point concerns the representativeness of the selected monitoring sites with respect to static features. While we agree that the overall number of wells and their spatial coverage may be sufficient to sample general hydrogeological characteristics, the wells included in this study were selected based on data availability and model performance considerations, not on the representativeness of their static environmental attributes. We did not explicitly assess, either a priori or a posteriori, whether the selected wells adequately cover the full range of static feature variability present in the broader dataset. Consequently, it cannot be excluded that the training and test sets underrepresent certain combinations or extremes of static environmental characteristics, which may limit the model’s ability to generalize to unseen wells. **We have revised the corresponding paragraph to clarify that well selection was driven by data availability and modeling considerations rather than by an explicit assessment of static feature representativeness (Line 232-239)**

The second point relates to the quality of the static environmental features themselves. We agree that the term “low-quality static features” may have been misleading. **In the revised manuscript, we have replaced this wording to avoid unintended implications. The corresponding paragraph has**

been revised to clarify that the limitations refer to uncertainties inherent to large-scale environmental datasets—such as coarse spatial resolution, interpolation, and indirect estimation of subsurface properties—rather than to erroneous or unusable data. We emphasize that these limitations reflect common challenges in large-scale hydrogeological datasets and do not invalidate the use of such features, but may reduce their explanatory power in a global modeling context (Line 225-231).

Taken together, our results suggest that both the representativeness of static feature distributions across wells and the inherent uncertainties and process relevance of large-scale environmental static features play an important role in determining model performance. A more systematic evaluation of static feature representativeness and the development of static attributes that more directly reflect hydrogeological processes would therefore be a valuable direction for future research, but was beyond the scope of the present study.

Overall, I don't quite understand how one could conclude that poor-quality static features have a greater impact on model performance than the integration strategy. Does this mean that all the used static features were poor quality? Or are these features considered poor quality compared to static features derived from time series, which are all meaningful? It will always be very difficult (not to say almost impossible sometimes) to have all at once high-quality, large-scale, high-resolution hydrogeological characteristics data that precisely accounts for spatial heterogeneity. Are we then reaching a major limitation to improve even more the generalization capabilities of Deep Learning models? In that case, would there be some particularly crucial environmental static features to focus on?

We thank the reviewer for this important and nuanced comment, which highlights the need to clarify the interpretation of our conclusions. Our statement that static feature characteristics may have a greater impact on model performance than the integration strategy should be understood in a relative, not absolute, sense.

We do not imply that all static features used in this study are inadequate, nor that integration strategies are unimportant. Rather, our results indicate that, within the range of integration strategies tested, differences in model performance were more strongly influenced by whether the static features were informative for the prediction task than by how these features were technically integrated into the model. In particular, time-series-derived static features—which directly summarize aspects of groundwater dynamics—proved to be consistently informative, whereas large-scale environmental static features are subject to inherent uncertainties related to spatial resolution, heterogeneity, and indirect estimation (see also our response to Comment 2.9).

We fully agree with the reviewer that obtaining spatially consistent and process-relevant hydrogeological characteristics at large scales is inherently challenging. However, we do not interpret this as a fundamental limitation of deep learning approaches. Rather, our findings suggest that future improvements in generalization performance are likely to depend more strongly on advances in the availability, resolution, and process relevance of static features than on further refinements of model architectures alone.

That said, identifying which environmental static attributes are particularly beneficial for large-scale groundwater prediction—such as features that more directly reflect aquifer properties or hydrological connectivity—remains an important and open direction for future research.

We have made several revisions to the manuscript, particularly in the Discussion and Conclusions sections, to clarify these aspects. In particular, we removed the term "poor-quality static features" and replaced it with more precise wording that reflects uncertainties and scale limitations inherent

to large-scale environmental datasets (Line 225-231). We further revised the corresponding paragraphs to distinguish more clearly between (i) uncertainties in static environmental data (Line 225-231) and (ii) the representativeness of well selection with respect to static feature distributions (Line 232-239). Finally, we reformulated the Conclusions to emphasize that the relative informativeness and process relevance of static features, rather than their "quality" in an absolute sense, exert a stronger influence on model performance than the specific integration strategy (Line 340-343).