Anonymous Referee #2

The manuscript describes a important groundwater level dataset for benchmarking machine learning models. I like the dataset because it is both important and timely for multiple disciplines, including hydrology, hydrogeology, and data science, etc. There is no doubt that it will be widely used for machine learning development, similar to the CAMELS datasets. I also like the manuscript because it is clear and concise. The dataset is well organized and follows the FAIR principles. I have few comments on the manuscript and dataset for the authors' consideration when they do some minor revisions.

- We would like to sincerely thank Reviewer 2 for the very positive and constructive feedback on our manuscript essd-2025-321. We greatly appreciate the reviewer's recognition of the dataset's scientific relevance, clarity, and adherence to FAIR principles.
- Our detailed responses are provided below and are organized directly under each individual comment for clarity.

Major comments.

Are the groundwater levels from shallow wells, deep wells, or a combination of both? Please clarify this information in the manuscript.

We thank the reviewer for this helpful remark. The dataset indeed comprises a combination of shallow and deep observation wells across Germany, covering both unconfined and confined aquifer conditions. Corresponding information is included in the metadata (fields *Depth*, *UpFilter*, *LoFilter*, *ScrLength*, *PreState*) and listed in Table 2. We have clarified this in the revised manuscript (Section 3.3, "Site-specific static data") by adding the following sentence:

> "The resulting dataset covers both shallow and deep monitoring wells under unconfined and confined conditions, as indicated by the available depth and pressure state metadata."

Although the imputation rate in the dataset is quite low, would it be good to include a synthetic test to evaluate the quality of the imputation (i.e., how well or poorly it performs)?

- We appreciate this thoughtful suggestion. The overall imputation rate in the dataset is indeed very low (mean 0.9 %, maximum 3.8 % during the test period), and missing values are typically short and evenly distributed. For this reason, we did not perform a separate synthetic test, as the influence of imputation on model performance was expected to be negligible. To verify this assumption, we compared benchmark model results based on the imputed dataset with an otherwise identical version where missing values were left unfilled. The resulting differences in RMSE, R^2, and Bias were marginal (see our response to Reviewer 1, Tab 1), confirming that the imputation had no relevant impact on predictive performance.
- Nevertheless, the imputation procedure was carefully designed to minimize potential bias: it leverages correlated neighboring wells as predictors, applies a block-wise strategy with temporal overlap to preserve seasonal dynamics, and explicitly flags all imputed values using the binary variable GWL_flag. This enables users to independently assess, validate, or re-implement the imputation process if desired.

The role of static covariates in machine learning model development (e.g., CNN or LSTM) is somewhat limited if data from all 3,207 wells are used for training, validation, and testing. I suggest the authors perform a spatiotemporal testing experiment — for example, hold out a group of wells entirely from the training set and use them only for testing — to evaluate whether the static covariates improve model performance.

- We appreciate this valuable suggestion and fully agree that spatial hold-out validation represents an important next step for assessing the contribution of static covariates to model generalization. However, the present study aims to establish the dataset and provide initial benchmark models that demonstrate its usability rather than to exhaustively explore all possible validation strategies.
- Our focus was therefore on transparent and reproducible baseline models using a consistent temporal split across all wells. Spatially independent evaluation experiments (such as leave-one-region-out or stratigraphic or clustered hold-out tests) are indeed planned for follow-up studies using the GEMS-GER dataset and are explicitly encouraged within its documentation.
- To reflect this point, we have added a short statement in the Conclusions section emphasizing that future work may include spatiotemporal validation experiments to systematically evaluate the role of static features in model transferability.

Minor comments.

Lines 91–101: Consider presenting the names in a table and including the number of wells associated with each. This would make it easier for readers to understand.

- We appreciate this helpful suggestion. The distribution of wells across the 16 German federal states is already visualized in Figure 2, which shows the number of wells per state at different preprocessing stages. For clarity, we slightly revised the text in Section 2.1 to explicitly refer to this figure when listing the data-providing authorities.

Line 111: The acronym LANUV has already been introduced earlier; the full name is not needed here.

- We thank the reviewer for pointing this out. We have removed the repeated full name of LANUV and now refer to the authority only by its acronym in this section.

Section 2.2.4: Please add one or two sentences explaining the possible reasons for the identified outliers.

- We thank the reviewer for this helpful suggestion. We have added two sentences at the end of Section 2.2.4 explaining that most outliers were related to technical or anthropogenic factors rather than natural groundwater dynamics (e.g., sensor malfunction or recalibration, data logger replacement, or short-term impacts from construction, pumping, or maintenance activities near the observation wells).

Line 186. You mean ‘density’ here, is that calculated by number of wells divided by area? or MHD1 has the highest number of wells?

- Yes, “density” refers to the number of monitoring wells per unit area (wells / km²) within each Major Hydrogeological District (MHD). We have clarified this in the text accordingly.

Lines 190–200: Are these variables listed in Table 2? If so, note that they could also serve as static covariates for machine learning model development.

- We deliberately did not include these dynamic indicators as additional static covariates in the dataset to avoid further increasing the already extensive set of static attributes (> 50 features). Moreover, they can easily be recomputed from the provided groundwater level and meteorological time series, allowing users to decide individually whether and how to integrate them into their own feature-engineering workflows.

- While such derived variables could potentially be used as static or quasi-static covariates in dedicated model experiments, the benchmark models presented here were intended as a compact and transparent “add-on” to illustrate the dataset’s usability rather than an exhaustive modeling exercise. We therefore intentionally kept the setup simple and encourage future studies to explore extended feature combinations.

Section 3: While the cleaned and quality-controlled groundwater-level time series are very useful, have the authors considered including the original (raw) dataset as well?

- We decided not to include the unprocessed raw data in the public release for several reasons. First, the raw groundwater-level series contain technical artifacts and implausible spikes that could easily lead to misinterpretations if used without the accompanying quality-control procedures. Second, the published dataset provides full transparency through the *GWL_flag* variable, which allows users to identify and exclude imputed values, effectively reconstructing the validated observations. Moreover, maintaining a second, unfiltered data version would create redundancy and complicate version control, potentially undermining reproducibility. Finally, the main purpose of GEMS-GER is to serve as a harmonized, quality-assured benchmark dataset rather than a raw data archive, in line with ESSD’s emphasis on usability and scientific robustness