Review of "EARLS: A runoff reconstruction dataset for Europe"

This paper presents an interesting contribution called EARLS. EARLS provides daily streamflow reconstructions (with uncertainty estimates) for over 10,000 European basins from 1953 to 2023, using a single LSTM-based rainfall–runoff model trained on data from more than 5,000 basins. The quality of the dataset is assessed through comparisons with held-out sets of basins and a qualitative evaluation of peak flows and flood timing between EARLS and results previously reported by Blöschl et al. EARLS is publicly available and serves as good example of how Deep Learning can be used to generate large-sample "datasets" with improved spatial and temporal coverage compared to existing observational datasets (e.g., EStreams in the case of Europe). The paper is well-written, and the analyses appear robust; thus, EARLS has the potential to be a valuable contribution to the hydrological community. In principle, I support the publication of this work without many changes. However, I also provide some related comments below, which I encourage the authors to consider to further strengthen the work and better justify its publication in ESSD.

Where does data end and model estimates start?

I agree with the opening paragraph that data are central to hydrological science. Observations are the primary reference for enabling hypothesis testing and advancing science (e.g., through studies utilizing LSTM models). However, the distinction between where "data" end and "model estimates" begin remains ambiguous. While it is true that observed hydrographs are not entirely free from assumptions (e.g., the construction of a rating curve), considering (LSTM) reconstructions as data on the same level is a significant leap. Are "reconstructions" (at least the parts where they provide estimates of data) not inherently limited in this regard? They offer gap-filled estimates dictated by the model, but they lack the inclusion of scientific testing that could uncover explicit new process knowledge beyond what is already encoded in the model.

The availability of streamflow estimates provided by EARLS is undoubtedly convenient, but is convenience what hydrological science needs? Scientific ideas should be tested using observations, and existing datasets (e.g., EStreams) provide such data. Gap-filled analyses should only be used when original data are insufficient, and even then, they are only useful if there is strong evidence that the gap-filling process accurately captures the specific hydrological behavior being tested. Consequently, a generic statement such as "Our results suggest that EARLS as a dataset is well-suited for large-sample hydrological studies in Europe" seems misplaced, as its suitability ultimately depends on the specific goals of the hydrological study.

Right now, these issue remains largely undiscussed, except for some comments in the conclusions. Is this the right place to briefly bring this up or should this be done earlier and more extensively?

How human-affected are catchments with <4 dams?

EStreams provides runoff estimates for catchments that have been screened for human impacts (fewer than four dams). However, in reality, there are over a million obstructions in European rivers (see: https://amber.international/european-barrier-atlas/). While not all of these obstructions significantly affect streamflow, the cutoff of four dams—regardless of catchment size—seems rather ad hoc. Why isnt a metric like dam density (e.g. per unit catchment area) or a similar approach used instead?

Comparisons are with the interesting studies of Bloschl et al. but are these the most informative comparisons?

Seasonal flooding is largely driven by climate conditions and has minimal connection to local landscape factors (e.g., see the geographical clusters in Fig. 7 and earlier studies attributing flood drivers). Could you devise a more challenging test that better demonstrates the utility of EARLS? For instance, you could quantify differences between observed and EARLS behaviors across several distinct and complementary hydrological signatures, ideally including a signature that is less spatially autocorrelated and that primarily reflects local differences between catchments rather than large-scale climatic gradients. While I understand the Bertola metric is used, this result is not presented spatially and also focuses on flooding. For example, I do not know whether EARLS is currently suitable for low-flow studies (or any other aspect of flow than annual high flows)

Is Fig 9 a reason to celebrate success, or does it highlight strong limitations?

I agree that some first-order inferences from Blöschl et al. are also apparent in EARLS. However, significant differences, including at larger scales, are evident (for example, many regions show opposing trends—consider Scandinavia—and I estimate that roughly 30% of the map displays inconsistent, opposing trends).

You state, you *"[…] argue that the results of our qualitative assessment show the merit our the* EARLS data for scientific inquiry and corroborate the quality of the simulations.."* However, to me, Fig. 9 suggests that I would be extremely hesitant to rely on EARLS for such analyses (given the inconsistencies highlighted above). Am I being too pessimistic?

*remove "the"

Why is the comparison in section 3.4/fig9 kept qualitative?

The comparison is currently qualitative, but transitioning to a quantitative analysis would be straightforward. Differences could be quantified and mapped at the pixel scale, while values across a continental scale could be compared using metrics such as mean absolute error, R^2, or other statistical measures. At present, the approach feels unnecessarily qualitative.

Please check lines 293-305.

(Maybe I found it challenging to follow your reasoning because I encountered this part towards the end of the review.) However, this section appears to be imprecisely formulated, leaving the reader too often to guess the intended meaning behind the words.

Data screening

I understand that the Estreams data has already undergone screening before publication, and you apply additional criteria.

However, it's important to note that some (artefactual) patterns likely persist in the data you're considering.

For instance, in the case of ES000454, it appears that patterns changed in 1966, coinciding with the construction of a dam. Additionally, the data for this station shows repeated nonzero values between 1973 and 1982.

I understand that data screening can be very time consuming, but did you also visually inspect the remaining hydrographs? This would be feasible for 5000 stations.

Detailed comments

- Fix the type on the first line of Eq. D3. (*m/2pi and not *m/365?)
- I understand that this paper is not about testing the trends in flood seasonality across Europe, but please note the methodological obscurity of Eq. D5: It is clear that one needs to take into account the circularity of the calendar year, and therefore, it seems logical that the variable *k* is introduced. However, for a trend in time, two processes can be more than half a year apart when considering their difference (and thus trend). Consider an example where flood peaks on average occur July 1st, but in one year the flood occurred early in the year (e.g. 15 March) and the subsequent year it was a late occurrence (e.g., 15 October). This suggests, for this pair, a trend towards later flooding, but the correction by *k* would qualify this a tendency towards earlier flooding. This could be solved by using a wrapping function, but applying this is probably outside the scope of the manuscript as your goal is to test how it compares to the results of bloschl.

- L6: are **<u>often</u>** well suited to provide predictions in ungauged basin" (as sometimes they clearly are not).
- L6-8: consider stating what the comparison shows, rather than that the comparison is made (mention the result/conclusion, not only the method)
- L90: out of curiosity: elevation is used, but elevation itself is not affecting hydrology, it are physical properties of the catchment that will correlate with elevation that do this. Do you have any idea which physical factors these are? (No changes required, just curious).
- L173: since NSE cannot really be compared between places I would not talk about "best and worst"but about "highest and lowest".
- L173: These values are not randomly distributed (Fig. 4): specify you talk about a distribution <u>in space (or geographically)</u>
- Caption Figure 8: "but show more detail." I understand they show more detail, but given the performance of the model I do not know if these "details" are right or artefacts, so I would be careful to emphasize this.
- L294: what do you really mean by: <u>virtual</u> variable?
- L296: what do you really mean by: is a new qualitative dimension?
- L296 "The job of the model is here to extract the information that the meteorological signals contain about the streamflow and act as a virtual sensor" is to me a vague statement.
- L297-299: "This is certainly not a replacement for real data — on the contrary, our ability to enable models in this way heavily depends on the availability of large amounts of diverse and highly qualitative data (Kratzert et al., 2024)." I do not fully follow you here.