Reviewer: Wouter Berghuijs

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 a 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.

Dear Wouter,

thank you for your thoughtful review. The review provided a lot of thoughts related to the big picture of the dataset itself and we highly appreciate this kind of feedback — even if it does not always lead to direct changes in the manuscript, we believe it improved the overall argument.

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.

We mostly agree with this part of the feedback and find the current discussion of our manuscript reflects it. For example, the statement that "streamflow measurements are of higher quality than simulations, despite being constructed too" is literally an argument that we make. We absolutely do not claim that our reconstructions are at the same level of accuracy as observations and will make sure to revise the discussion so that this becomes crystal clear to all readers.

There are, however, two small aspects that we would like to clarify: Firstly, data does not end where model estimates begin. Estimates are, by definition, data. Secondly, the property that one can only extract the information that is contained in the data is a fundamental property of information in data (see for example Chpt. 2.8 in Cover and Thomas, 2006). As such, it is shared by all forms of data irrespective of their provenance and independently how much scientists try to analyse or process the data.

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.

We agree with this reflection in spirit. Observations and experiments are the ultimate arbiter in science. At the same time science is much more than a mere collection of ideas, factoids, and data points of observables. It is conjecture; it is exploration; it is the weaving together of ideas, understanding, and empirical reality; it is the generation, management, and updating of knowledge; it is much more — and in all of this in-silico results can (and, as a matter of fact did already!) help.

Hence, we would (a) like to avoid dictating to scientists what they require or not, and (b) argue that the argumentation is formulated too broadly. Many sciences already make extensive use of large-scale simulation datasets. Perhaps the most famous example from earth science is climate science (but there are many more). We are not aware of any distinguishing factor that would hinder hydrologists to also account for evidence from simulations—keeping in mind that the resulting insights stem from simulations, as discussed in the answer above (and, as a matter of fact, in practice they do).

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?

We appreciate the concern and recommendations. However, our manuscript extensively discusses these issues; particularly we would say that even more than any other comparable dataset published on ESSD. For instance, we show that EARLS can be used to investigate long-term streamflow trends in similar ways as done in previous studies. We therefore believe that a more extensive discussion would not improve the manuscript.

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 isn't a metric like dam density (e.g. per unit catchment area) or a similar approach used instead?

We appreciate your concern, and indeed from a modelling perspective this is slightly arbitrary. We could also train our model on the whole dataset without pre-filtering

and derive predictions/evaluations. However, we decided to do some data curation because it is associated with better training behavior and performance of machine learning models. And, questions about how and how much one should curate the data are always a matter of convention and taste. There are infinitely many criteria to choose how to filter the data and our model does not strictly require perfectly unaffected streamflow. The one proposed here is one realization of these criteria.

That said, we like the proposed idea per se. As we write in our manuscript, EARLS is meant as a living dataset. Hence, we happily support other versions that use different delineation criterions for future expansions of the dataset. As such, we are happy to include such criteria first into a new EStreams version and then build a new EARLS version out of it. For the revised version of the manuscript we will mention this idea as future work.

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?

We argue that these comparisons are suited to show the utility of EARLS. They certainly go well beyond all published hydrological datasets that we are aware of.

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)

It is not within the scope of our role as data creators to develop demonstrations for the wide array of potential hydrological applications. Given the diversity and breadth of possible uses for the dataset—many of which we may not yet be aware of—we

focus on ensuring the dataset is robust and well-documented, allowing researchers to tailor its application to their specific needs.

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?

We appreciate your concern, but we have the impression that this question arises from a misunderstanding of what the data from Blöschl et al. (2017, 2019) constitute. It is important to realize that there is no ground-truth to compare with in the first place. Consider the following: The maps from Blöschl et al. (2017) are constructed with the help of a small number of highly processed point sources, which are rasterized over large areas without observation by using a "gap filling" approach based on kriging. When we say gap filling, we use the terminology by the reviewer here, but we note that this filling is absolutely model based. Our reproductions with EARLS use the same gap filling mechanism, but the (naive) kriging is supported by a much denser distribution of many more support points. Since it was mentioned in the comment we take Scandinavia as an example (see Figure A1 below): If we consider a box estimate of the measurement station from Blöschl et al. (i.e., points that lie within a certain box of latitudes and longitudes) we get approximately 300 (here we rounded up in the decimals and neglect that they filter out specific stations for some of their analyses). In comparison, EARLS contains around 2500 simulated stations in the same area (here we rounded down in the hundreds). In that sense, 30% difference might not be bad. Either way, there is no ground truth to compare with. Since we can see where the reviewer came from we will add this discussion to the appendix of the revised manuscript.

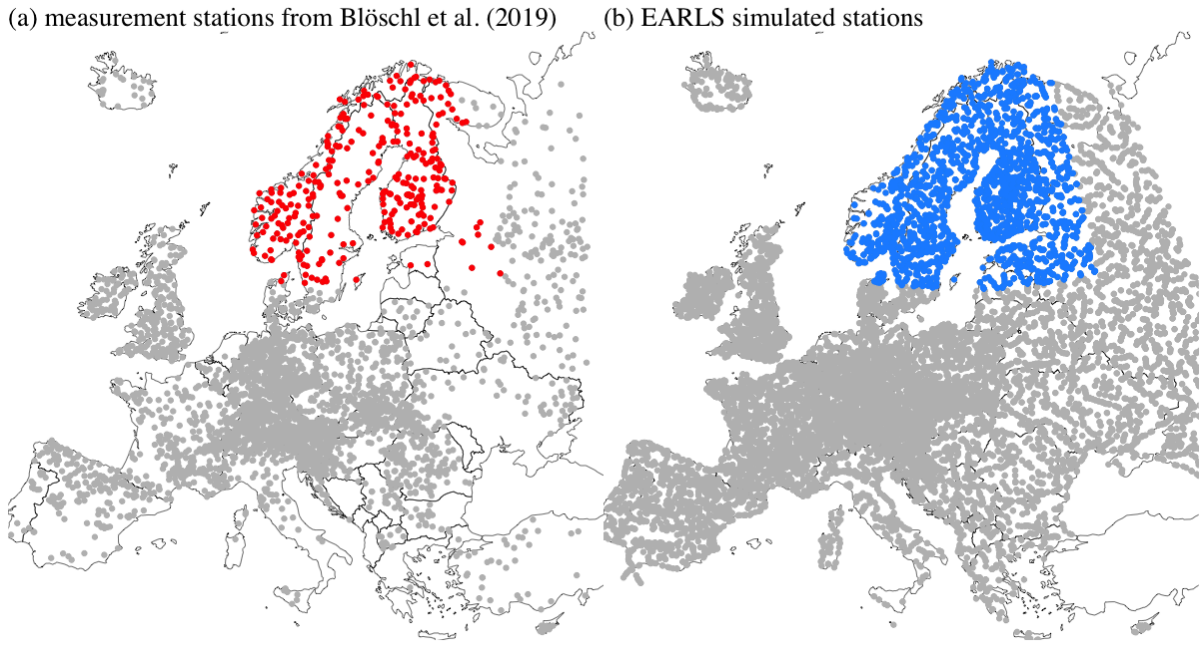

Figure A1. Comparison of the density of support points for the kriging interpolation exercise. Plot (a) shows the reference points from Bloesch et al. (2019), plot (b) the simulated station from EARLS. The colored areas in both (a) and (b) refer to the same subselection of the data based on boxing a given latitude and longitude.

*remove "the"

Thank you.

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, R2, or other statistical measures. At present, the approach feels unnecessarily qualitative.

This question probably goes back to the misunderstanding that caused the previous question. In short, this section is kept qualitative, since our goal is to show similarities in patterns. None of the proposed metrics capture that aspect. Not only that, using such metrics in this context would easily mislead readers into thinking that the Blöschl et al. (2017, 2019) figures are a form of ground truth. They are not. In the revised manuscript we will make sure that this point is clear to the readers.

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.

Thank you. We will go over it again.

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.

We did visually inspect many hydrographs by randomly sampling different time-spans from different years.

Additionally, similarly to EARLS, EStreams is expected to be constantly updated, and we expect to have a new version with such visual inspection performed to all 17,000+ stations in the near future. As such, we will be happy to include such criteria first into a new EStreams version and then build a new EARLS version out of it—after the publication of the first version is finished. We will refer to this as future work in the revised manuscript.

Detailed comments

- Fix the type on the first line of Eq. D3. (*m/2pi and not *m/365?)

Thank you for pointing this out. We will correct this.

- 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 Blöschl.

Even if there are peculiarities in the approach we would like to keep them in our reproductions to make the comparison variable. We will, however, address this concern in the revised manuscript and mention that the computation is not necessarily intuitive.

As a side note: If we assume that the two events are the same, both the numerator and denominator would be negative. Hence, the tendency would be towards later flooding not earlier ones. Either way, as the reviewer points out himself, it has little to do with our comparison.

- L6: are **often** well suited to provide predictions in ungauged basin" (as sometimes they clearly are not).

Ok.

- L6-8: consider stating what the comparison shows, rather than that the comparison is made (mention the result/conclusion, not only the method)

We will revise this part as proposed.

- 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).

No. Unfortunately, we do not (in general the factors do not just have to be physical, e.g., they can also be biological or anthropogenic).

- L173: since NSE cannot really be compared between places I would not talk

The goal here is to provide a ballpark number. We will make sure to emphasise that.

- about “best and worst”but about “highest and lowest”.

Thank you. We will correct this.

- L173: These values are not randomly distributed (Fig. 4): specify you talk about a distribution in space (or geographically)

We will clarify by rephrasing to “These values are not randomly distributed in space (Fig. 4)”

- 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.

The details are actually important in that they show that we ingest additional information via the model at the simulation and do not just naively smooth between scarce observations. We will make sure to emphasise this in the revised manuscript.

- L294: what do you really mean by: virtual variable?

We mean it in the sense of Beven et al. (2012).

- L296: what do you really mean by: is a new qualitative dimension?

With this we want to express that we still argue that simulations are qualitatively different from the streamflow “observations”. Basically, the sentence makes the same argument as the first question/comment by the reviewer. We will make sure that this aspect will be more evident in the revised form of the manuscript.

- 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.

We will rephrase this sentence to make it more clear. In the revised version we will write:

In our case, the model acts as a virtual sensor. It extracts the information that the meteorological signals contain about the streamflow. The ungauged EARLS basins do constitute a layer of information that one would not obtain from using streamflow observations only.

- 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.

With this sentence we want to express that observational data is and will remain the most important factor for hydrological inquiry. We will formulate this clearer in the revised version of the manuscript.