General comments

The manuscript delivers an interesting addition to the current surge of machine-learning in hydrological modelling by extending the application of LSTMs from pure streamflow modeling to actual forecasting. To do so, they ingest the output of physical forecasting systems as input to an LSTM. The main result, that LSTM outperforms the other approaches, is not surprising as the LSTM merely acts as a bias correction algorithm with many more degrees of freedom. Nevertheless, this is still a relevant finding that should be disseminated throughout the hydrological community. The manuscript is well-structured and comprehensible, particularly the introduction and methods parts are very comprehensive yet concise. Intuitive measures of model performance, a solid discussion of the error metrics as used in the study, a comprehensible discussion on the nature and choice of datasets as well as informative plots act as a solid foundation for the reader to follow along w.r.t to the methodological execution and its results. However, towards the end, the discussion and conclusion do miss out to put the work and the results into a broader perspective e.g. by contrasting it against ongoing machine-learning research, its limitations or future directions for hybrid modeling (see detailed comments below).
We thank the reviewer for giving his time to provide a detailed review of our manuscript. We respond to his points individually below, in red.

Scientific/Specific comments

1. Please elaborate on the different types of "Hybrid" models/forecasts that are possible. In ML literature, there are current advances, termed "hybrid models", of including and solving differential equations inside the NN, promising the best of both worlds (high accuracy while keeping interpretability/robustness to out-of-distribution cases). These developments should be listed as future directions of research and the approach of this manuscript should be contrasted against these new development in the introduction.

Rackauckas, Christopher, et al. "Universal differential equations for scientific machine learning." arXiv preprint arXiv:2001.04385 (2020).

Raissi, Maziar, Alireza Yazdani, and George Em Karniadakis. "Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations." Science 367.6481 (2020): 1026-1030. APA
We agree with this comment and will add these discussion points and references to the revised introduction so that it also covers the limitations and future directions of hybrid modelling in hydrology. Thank you for the suggested references.

2. As the title suggests, the LSTM-approach is set up to "boost" the forecasts of IFS. As the LSTM receives Era5/IFS streamflow estimates, I personally would rather view it as a more sophisticated bias correction (more parameters, less constraints) than a

separate modeling approach. I believe that the manuscript would benefit if this was put into perspective in the discussion part. Also, the possibility of using a simple statistical/ML model with less training effort, e.g. a simple linear model or RandomForest should at least be mentioned as an alternative.
We disagree with this. Although the boundary is vague and ML statistical correction, hybrid modelling, and pure ML models exist on a spectrum, the method implemented here would still work even if GloFAS were removed from the input (though perhaps with reduced skill). NWP variables are used as input, and so the LSTM is in some way replicating the hydrological processes, as in the Kratzert papers cited throughout.

3. Especially when taking into account that 7/10 catchments are known to be too small for the raw GloFAS to perform well, it is obvious that a statistical bias correction outperforms the raw forecast more the more degrees of freedom it is given. The "unfairness" of the comparison of raw GLOFAS vs. LSTM (world-wide simulation at 0.1° resolution vs. local model) should be highlighted in the introduction and discussion sections.
Yes – this is of course part of the motivation for trialling an LSTM in the first place. We will add this caveat to the revised introduction and discussions.

4. It would be beneficial to include and elaborate a bit further the motivation behind the two bias correction algorithms in the introduction to 4.2., i.e. quantile (remap values to reduce systematic bias) and spatial (inherent spatial bias in GLOFIS-ERA5(?)).
Quantile mapping is a fairly standard practice in hydrological bias correction (e.g. Thrasher *et al*., 2012). The spatial mapping provides an additional layer of bias correction, accounting for the fact that consistent spatiotemporal biases in hydrometeorological fields such as precipitation (should they exist), will result in consistent upstream/downstream biases in streamflow – information that can be used to improve forecasts. We will expand the revised methods section to include this motivation.

Thrasher, B., Maurer, E. P., McKellar, C., & Duffy, P. B. (2012). Bias correcting climate model simulated daily temperature extremes with quantile mapping. Hydrology and Earth System Sciences, 16(9), 3309-3314.

Also, the motivation for the two final steps should be explained and justified in greater detail. What do you mean by different climatologies, why split 3/4 vs. 1/4? Why do you shift the forecasts that have ben quantile mapped-once more?
This is a fair question and has also been raised by reviewer 2. GloFAS-ERA5 and GloFAS forecasts have different climatologies because they take meteorological input from different sources (ERA5 and IFS forecasts respectively). ERA5 and IFS themselves have different climatologies because, although they share the same driving model, ERA5 is a reanalysis and is nudged towards observations, whereas

IFS forecasts aren't. We realised at the beginning of the operational phase that the differences between GloFAS and GloFAS-ERA5 were leading to an overenthusiastic bias correction, and so damped it using the weights given. This point was also raised by reviewer 2, and we will expand the revised methods section to include this explanation.

Also the fact that the bias correction has been newly developed (mentioned in conclusions) should be placed in the respective chapter in the methods part.
Yes – we will mention this in the revised methods section.

5. You argue that you used reanalysis data during train+test to make the results reproducible for potential users. But are the operational forecasts using IFS still reproducible? If not, it would be beneficial to provide the respective data on zenodo or a similar platform.
The IFS data used for these forecasts is freely available from the MARS web service hosted by ECMWF. The full streamflow forecasts themselves are available on the project GitHub page.

6. Generally, the manuscript misses to give detailled information on the training process. I would advise to include loss curve(s) (loss vs. epoch) of test and train. This is the common way to present information to see whether training was successful. Also, train and test error metrics should be provided to give an intuition whether under- or overfitting might have happened. The same applies to the bias correction, here the reader is not provided with any information on the optimization procedure or performance, even though there is a risk of overfitting. Similarly, information on the loss function, training hyperparameters (dropout, decay, learning rate, recurrent activation etc.) should be listed in appendix or refer to repository. To this end, the training process could have been performed more thoroughly: Hyperparameter-Tuning usually involves searching over model/training parameters as well as model configurations (n hidden layers, nodes etc.), not only epochs. The latter is usually less relevant. Ideally, hyperparameter tuning would be executed in a cross-validation set-up. To that respect, please elaborate what "tuned using sensitivity tests" refers to (l.240).
We agree that we have provided too little detail on the LSTM training process – although we would like to note that as the code is open access, a sufficiently interested reader could find the finer details there. However, in response to this suggestion and in keeping with ML literature conventions, we are happy to expand the methods section to include more information on (a) the final choice of hyperparameters, (b) how we arrived at these choices, and (c) how we avoided overfitting.
With the bias correction, which is ultimately a linear – if somewhat multivariate – model, we can be fairly sure that overfitting has not occurred for two reasons. Firstly, the number of coefficients sought is several orders of magnitude less than the number of training data points. Secondly, the validation – for which we used

data set aside from the training/fitting process – suggested a very good fit. We will add these points to the revised methodology.

7. The conclusion is (too) detailled on the technical side (n of skilful vs. non-skilful results, KGE vs. NSE) but misses to provide the most relevant point in 2-3 comprehensive sentences: Where exactly does LSTM perform best/worst (seasonal, diurnal, altitude etc.), Where are its limitations? Both the discussion and conclusion miss to put the results into a broader perspective: How could the LSTM be improved other than switching to Convolutional LSTM-layers? What are future research directions? Maybe some points to consider here:
We agree that some more detailed discussion on limitations and future research areas for ML in hydrology would be useful, as well as a synthesis on overall performance. We respond to the suggestions individually below.

- Limitations of LSTMs: Some limitations are fairly well-acknowledged by now (parallelization, long-range dependencies) so that LSTMs are not the go-to model for sequential data anymore. Thus, outlook on new developments like GRUs and, most importantly attention-based models (e.g. transformers) should be included
  Although LSTMs do have some limitations, they have been shown to be an incredibly powerful tool when it comes to modelling hydrological systems (e.g. Gauch *et al*, 2021). Of course, as ML models become increasingly sophisticated, there is little doubt that novel architectures/technique will perform such tasks much better, but there is very little (if any) published research on that yet. We will add a short discussion of these new developments in general sequential modelling to the revised discussion.

  Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrology and Earth System Sciences, 25(4), 2045-2062.
- Limitations of ML generally: Generalisation when provided with out-of-distribution data, e.g. due to system changes (climate change), lack of interpretability
  These are indeed often cited as shortcomings of ML. However, as shown in Kratzert *et al*. (2019), LSTMs trained across multiple basins actually perform well on previously unseen (even ungauged) basins that might typically be considered out-of-distribution. Similarly, such LSTMs can capture extreme values of streamflow driven by out-of-distribution extreme precipitation events (Frame *et al*., 2021). Taken together, these results suggest that such LSTMs are capable of capturing the underlying hydrological relationships that connect precipitation, runoff, and streamflow. Since climate change doesn't affect these physical relationships (only the magnitudes of the inputs), sufficiently advanced LSTMs should be largely immune. That said, other types

of changes, e.g. increasing urbanisation, can affect the underlying relationships, and would degrade the skill of the LSTM – though this would of course happen in any other type of hydrological model if not updated. Although recent work has shown potential for interpretability in streamflow LSTMs (e.g. using attention theory; Li *et al*., 2021), we appreciate this is still generally a weakness compared to physics-based models – although in not having to rely on prescribed relationships, ML products can potentially learn new ones.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12), 11344-11354.

Frame, J., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., ... & Nearing, G. S. (2021). Deep learning rainfall-runoff predictions of extreme events. Hydrology and Earth System Sciences Discussions, 1-20.

Li, W., Kiaghadi, A., & Dawson, C. (2021). High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. Neural Computing and Applications, 33(4), 1261-1278.

- Greater picture:
    - Comparison to other hybrid modelling approaches in ML (see above)
      Yes – as with your first major comment, we are happy to expand the discussion to include greater coverage of hybrid modelling approaches.
    - How far away are we from entirely ML-based forecasts, given considerable advances in ML-based climatological forecasts? Should science still focus on improving relatively coarse physical model like ERA5 or rather explore ML-based bias correction at a large spatial scale?
      It is natural to speculate on this, but one may as well ask "why should we continue improving cars when we have planes?" Ultimately, physics-based and ML-based hydrological models should both continue to be improved since, although they can be used for the same purpose (e.g. forecasting, as in this paper), they have different strengths and weaknesses – as discussed above –and will thus continue to be suitable for different applications.

- Reflect on the risk of overfitting in bias correction + lstms, including the fact that, ideally, one would have to estimate the bias of IFS w.r.t to ERA5.

Yes – this was one of our biggest concerns, as we stated in the original discussion: “Similarly, the LSTM was trained on ERA5 data, but then ingested IFS output when run operationally. Although the two products will share some biases, they will inevitably be larger in IFS (the forecast) than ERA5 (the reanalysis), resulting in errors that propagate non-linearly through the LSTM. Originally, we chose to train the model on ERA5 so that our methods were reproducible, but there is no reason other forecasters should be bound by this desire. The optimal strategy is to train the LSTM on IFS hindcasts – though this would require careful adjustment of the architecture to account for different lead times. Similarly, such an approach must be careful of changing hindcast model versions.”

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Technical Corrections

- The abstract could be shortened
  We appreciate that at ~400 words, the abstract is a little on the longer side. However, the paper covers quite a wide variety of work and new methodology, so it isn’t necessarily obvious how it could be shortened without loss of impact. We are happy to take guidance on this.
- l. 85, add bracket
  Thanks – we have fixed this.
- l. 110: "interesting challenge" is a subjective statement, rephrasing is advised
  We will remove “interesting” in the revised manuscript.
- l. 252 wether A extending
  Sorry – not really sure what the suggestion is here.
- l.262 remove note
  We have removed this.
- l. 281: Is it 6 or 7 catchments that are skilful? 5.1.1. lists 6/10 that are skilful, here it is "still" 7/10?
  Good spot – we have edited this in the revised manuscript so that it now says: “Following bias-correction, GloFAS-ERA5 is now skilful at seven stations and highly skilful at four.”
- l. 312 remove (CHECK)
  Thanks – we have removed this.
- caption fig. 9: "black" is actually grey
  Thanks for spotting this. We have made this correction.
- Please elaborate what the control member in GLOFAS/IFS (ll. 129;467) and the "ensemble member" of the LSTM are (fig. 5)

Details on the LSTM ensemble are already given in methods in Section 4.3 (L245-250 in the submitted manuscript). Control members of GloFAS/IFS are simply unperturbed members of their respective ensembles, which we will clarify in the revised manuscript.

- The detailled description of and motivation behind the choice of training, testing and operational timespans could be placed elsewhere than in 4.2.1. Possibly best at beginning of chapter 4. Please also make it clear that LSTM and bias correction use the same time spans.
  Following this comment and a similar one by reviewer 2, we intend to provide a synthesis of the full methods/workflow for quick reference at the beginning of our revised section 4.

Potential changes for readability:

- 4.3. Include input features as table, not in text
  We agree that this will improve readability and will put the variables into a table (either in the methods section or an appendix).
- Present accuracies (aka skilfulness) of the three models as a table
  We will look at some different ways of representing these data in a table and include one in the revised manuscript if suitable.