Reviewer comments – Reply – Changes to the manuscript

Reviewer #1: Journal: Geophysical Model Development (GMD)

Please note that all following line references correspond to the article without track changes.

General Comments: The authors present a machine learning approach to forecasting river runoff from weather data using convolutional long-short-term memory neural networks. They present convincing evidence that the utilized ml model shows results of equal quality as its training data. At the same time the ml method offers faster processing speeds and thus an easier direct integration into regional climate models. With this approach, they present a scientifically significant and qualitative contribution to the integration of river runoff forecasting into climate models. While the manuscript shows great potential, I think that it requires minor revisions. My comments are listed below.

Minor Comments: As I do not possess a deeper understanding of the river runoff modelling and come from the machine learning side, I will limit my comments mostly to the technical aspects.

First, to me it does not become clear exactly how well your training data performs in comparison to other state-of-the-art models. I understand that your ConvLSTM is able to reproduce its training data's quality but I'm not fully able to grasp the strengths and weaknesses of the utilized training model, which I can assumed are transferred to the ML model. It would be helpful to extend the technical details section or the model section by a short description of the training data and especially its strengths and weaknesses compared to other possible runoff forecasting models. Although I see that the point of the paper is more the proof that is able to reproduce a state-of-the-art river runoff forecasting and not the exact strenghts and weaknesses of the utilized training data, it would help give perspective to the strengths of your method

We agree that this is a good idea and added the following text to the paper.

L168-170: To this point, no comparable long-term dataset with daily resolution was available. In other studies multiple datasets have been merged, but offer only monthly resolution (see e.g. Figure 3 \citep{Meier2019}) .

For example, your training data seems to present a bias compared to observational data (Figure 7b), which the network reproduces.

You are right for pointing this out. However, this bias is likely caused by other factors than the river runoff. We added the following part for clarification:

L271-281: It should be noted that the discrepancies between the simulated salinity and the observed values at BY15 are not directly linked to the performance of the ConvLSTM river runoff model. Instead, it is attributed to the MOM5 ocean model's representation of physical processes, particularly the treatment of mixing, advection, and stratification in the Baltic Sea.

Several factors may contribute to this discrepancy. The Baltic Sea is known for its strong vertical stratification due to the input of freshwater from rivers. The MOM5 model uses the K-profile parameterization (KPP) scheme for turbulence, which may not perfectly resolve small-scale mixing processes and vertical salinity gradients. This can result in an overestimation of salinity variability at the surface. Moreover, while the MOM5 model captures the large-scale dynamics of the Baltic Sea, the lateral transport of saltwater from the Skagerrak into the central Baltic Sea may not be perfectly represented. This can introduce variability in surface and bottom salinity that are not observed in reality. However, all in all, the long-term trends and larger salinity changes are accurately captured, indicating the model's robustness in predicting high-frequency and low-frequency variations.

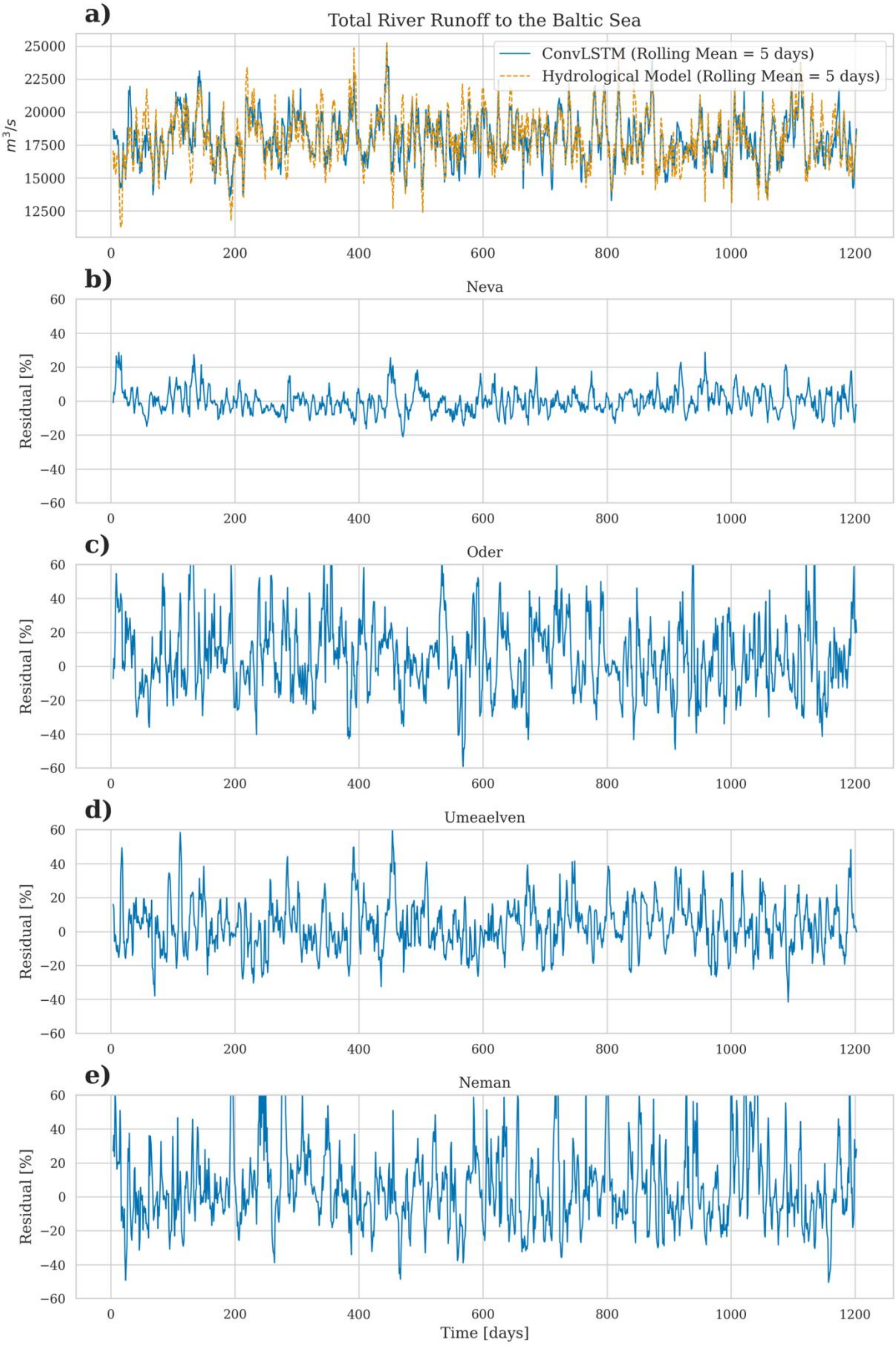
In connection to that, you describe that you utilize the time period from 1979 to 2011, because they are not bias corrected. As a bias correction seems to be usually conducted, I would like to know if that can be similarly performed on the ConvLSTM outputs.

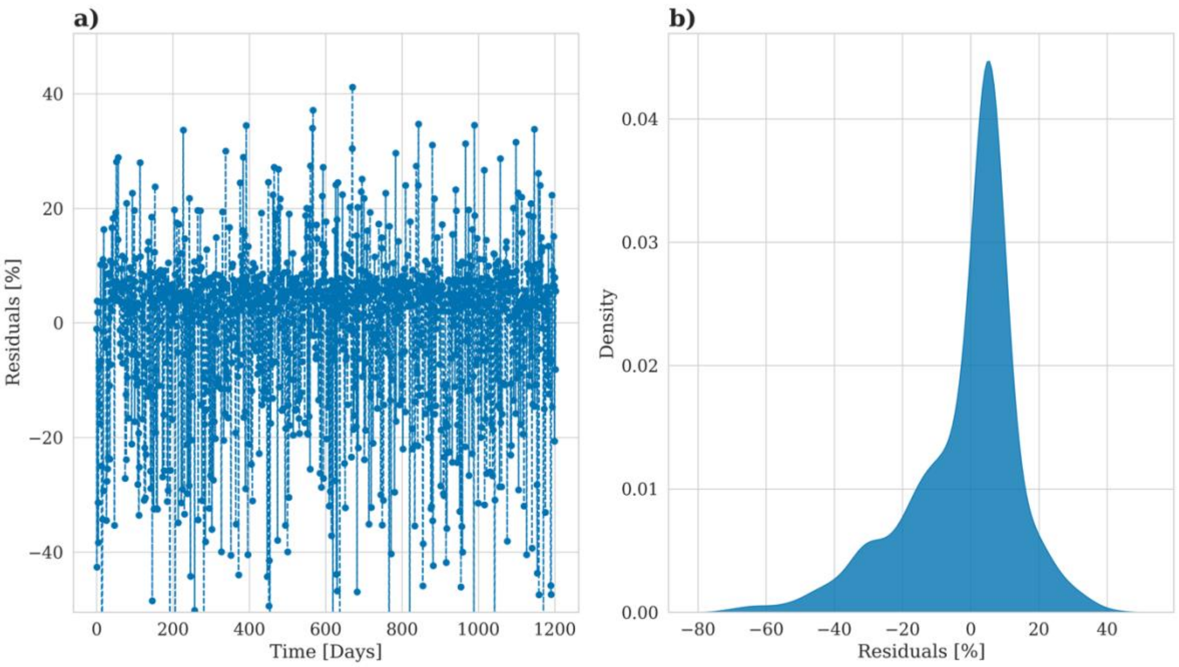
Yes, in principle this bias correction is possible, likely by additional scaling as post-processing. On the other hand, some of these bias corrections have been performed due to reasons like changes to the water-management or a new dam. This is why we chose to use a period where none of the bias corrections have been applied.

Connected to that, have you tried to train the ConvLSTM on any other runoff models? Training for 400 epochs on daily training data from 32 years is a lot of training input. Just out of interest, have you tried training on less data and how does the performance of the ConvLSTM differ? I would guess, that not all hydrological models provide such a comprehensive dataset. Could you thus comment on how easy it would be to extend this method to other runoff prediction models and how much training data would be required.

We did some preliminary test with other measurement data, but the results we mixed. Rivers that have a good temporal coverage with no changes in the way the data was measured performed well, while other rivers where structural changes were done performed poorly. Hence, we assume that the good performance of the ConvLSTM is also based on the good data availability.

For this review we also used only 10 years of training data (E-HYPE). In its current configuration the model’s performance is worse when only 1/3 of the trainings data is used.


a)
Total River Runoff to the Baltic Sea
ConvLSTM (Rolling Mean = 5 days)
Hydrological Model (Rolling Mean = 5 days)
m^3/s
25000
22500
20000
17500
15000
12500
0
200
400
600
800
1000
1200
b)
Neva
Residual [%]
60
40
20
0
−20
−40
−60
c)
Oder
d)
Umeaelven
e)
Neman
Time [days]



I would also be interested, if all ocean/regional climate models are able to utilize runoff predictions from similar sources or if they require their own in-model consistent runoff forcing. Because, if other climate models would require the ConvLSTM to be trained on different runoff predictions, it would significantly limit this method's applicability if that runoff model would be required to possess such a comprehensive training dataset as the EHYPE model presented in your study.

In general, the runoff data (all individual rivers) are mapped onto the ocean grid as a mass flux. While the grids may differ, the procedure is similar across all major ocean models.

Additionally, I would be interested out of curiosity how many timesteps are necessary for the LSTM to significantly improve the CNN output. Have you tried training with significantly less than 30 timesteps? What was your reasoning behind choosing these 30 days? Or was it just based on model performance/loss functions?

Based on your suggestion we performed several sensitivity tests of the hyper parameters. The model performance turns out to be relatively robust, even for shorter time steps (10 days). However, we still decided to use longer time scales, as we assume that longer input sizes increase the stability of the model needed for long-term climate simulations.

L209-213 The model's performance can be described as relatively robust when changing the set of hyper parameters (see Figure \ref{fig-Supp3}. Interestingly, also shorter input sizes of 10 days perform really well. However, we still decided to use longer time scales, as we assume that longer input sizes increase the stability of the model needed for long-term climate simulations.

Finally, you claim that "While the initial training of the model requires substantial computational resources, it remains significantly less intensive than running comprehensive hydrological models" (Page 17). Could you give an estimate on how big this“significant” reduction of computational resources is? Because in the end this time saving is the important improvement of your method compared to other numerical prediction systems/models.

We agree that this information is useful and added it to the text. Our model generates one year of daily river runoff in roughly 10 seconds. The runtime of a hydrological model (personal communication with Stefan Hagemann (Dr. Stefan Hagemann, Regional Land and Atmosphere Modeling, Head of Department), with a similar resolution varies between 5-15 minutes per year.

This results in a speed up in the range of factor 30-90.

L309-310: The achieved speedup (depending on the complexity of the hydrological model) is within the range of 30 to 90 times faster.

In general I felt the content of the paper was novel and the method would be of interest to others in the field, but some details should be explained further or lack a bit of background information.

Thank you very much and also thank you for taking the time to review the article.