Replies to https://doi.org/10.5194/egusphere-2025-4244-RC1 for the preprint https://doi.org/10.5194/egusphere-2025-4244, by Saint Fleur et al. (2025)

We would like to thank the reviewer for these very constructive and stimulating comments which will help improving the initial version of the manuscript. We provide replies and explanations to the reviewer's questions in the following.

Comment 2.1

The authors indicate that they are benchmarking against Kratzert2019 and Newman2017, but they have different training / testing periods as the benchmarks. In section 2.1 you indicate that you are training from 1989-2006 and testing in 2006-2008. However, Kratzert used 1999-2008 for training and 1989-1999 for testing. This means you are training much more than them, and testing in only a 2-year period. Why this difference?

Moreover, you are also indicating that you re-simulated the 1989-2008 period with the pretrained models from Kratzert and Newman, but again, you are not respecting the training/testing split that the authors did in their original study.

Using benchmarks is an extremely valuable technique, because it automatically places your method in existing literature, however the conditions of the original studies need to be respected. One should, if possible, adapt the new experiment to the existing benchmark, otherwise it does not make any sense to do a benchmark.

We thank the reviewer for this important comment. Using training and testing periods that differ from those in previous studies complicates direct comparisons. As pointed by the reviewer, the selected periods also imply that all approaches in our manuscript were not evaluated under identical conditions. Our work is built on Kratzert et al. (2019) and Newman et al. (2017), those reference models were trained and tested using matched periods. In our current setup, the 2006-2008 period used to evaluate the data assimilation (DA) strategies lies within the training period of the reference LSTM and SAC-SMA models. This choice inflates the apparent performance of the baseline models and likely disadvantages the DA approaches.

In the revised manuscript, we will align the training/testing split across all the approaches, using 1999-2008 for training and 1989-1999 for testing, ensuring consistent comparison.

Also, why are you using only two years of testing?

Our initial choice was to maximize the size of the training set, following common practice in AI-based studies, at the expense of a shorter testing period. Although a smaller testing set may increase the influence of sampling variability on the results, this is partially compensated

by the large number of basins in the test data set. Nevertheless, in light of the previous comment, we will revise the training/testing split in the updated version of the manuscript.

Lastly, you should benchmark your study against other studies that used data assimilation methods. In CAMELS US there is the study of Nearing (2022), Feng (2020) or more recently Yang (2025).

We thank the reviewer for bringing these studies to our attention. References to Nearing (2022), Feng (2020) and Yang (2025) will be considered in the revised manuscript. Pending further verification, some of the approaches proposed in Nearing (2022), as well as other relevant studies, will be included in our comparative analysis.

Comment 2.3

In section 2.3 you indicate that a specific model should be calibrated for each lead time, and that the alternative is inefficient. But this is not true. The LSTM has a linear layer at the end that transforms the hidden states to discharge.

Your model is a simple extension of this, but instead of a linear layer that go from hidden states to discharge, you have a feed forward neural network (so a couple of linear layers), in which, besides the simulated discharge you concatenate some past meteorological variables and discharge. I do not have anything against the simplicity of the model, because if it is simple and it works then great, but there are multiple things that need to be considered.

You can run the LSTM as a seq-seq model, roll it over the forecast horizon (so if you are predicting 7 days forecast just do seq-7) and in each step concatenate the hidden states with whatever you want and pass it through the feed forward neural network. This way you have a consistent and generalizable model that does not require a different embedding for each case.

The alternative we had in mind is the recursive forecasting approach. The reviewer is correct, the seq-to-seq architecture is another alternative which also introduces additional challenges, particularly the risk of obtaining intermediate performance across lead times. Training may become more complex and prone to sub-optimal solutions for individual lead times. Nonetheless, as shown in Nearing (2022), seq-to-seq models remain an interesting option in our perspective.

Also why are you evaluating only on days 1, 3 and 7? Why not all seven days?

This choice was for time saving reasons, the results for the intermediate lead times are expected to evolve monotonically. However, to avoid ambiguity, results for the complete range of the 7 days lead time will be included in the revised version.

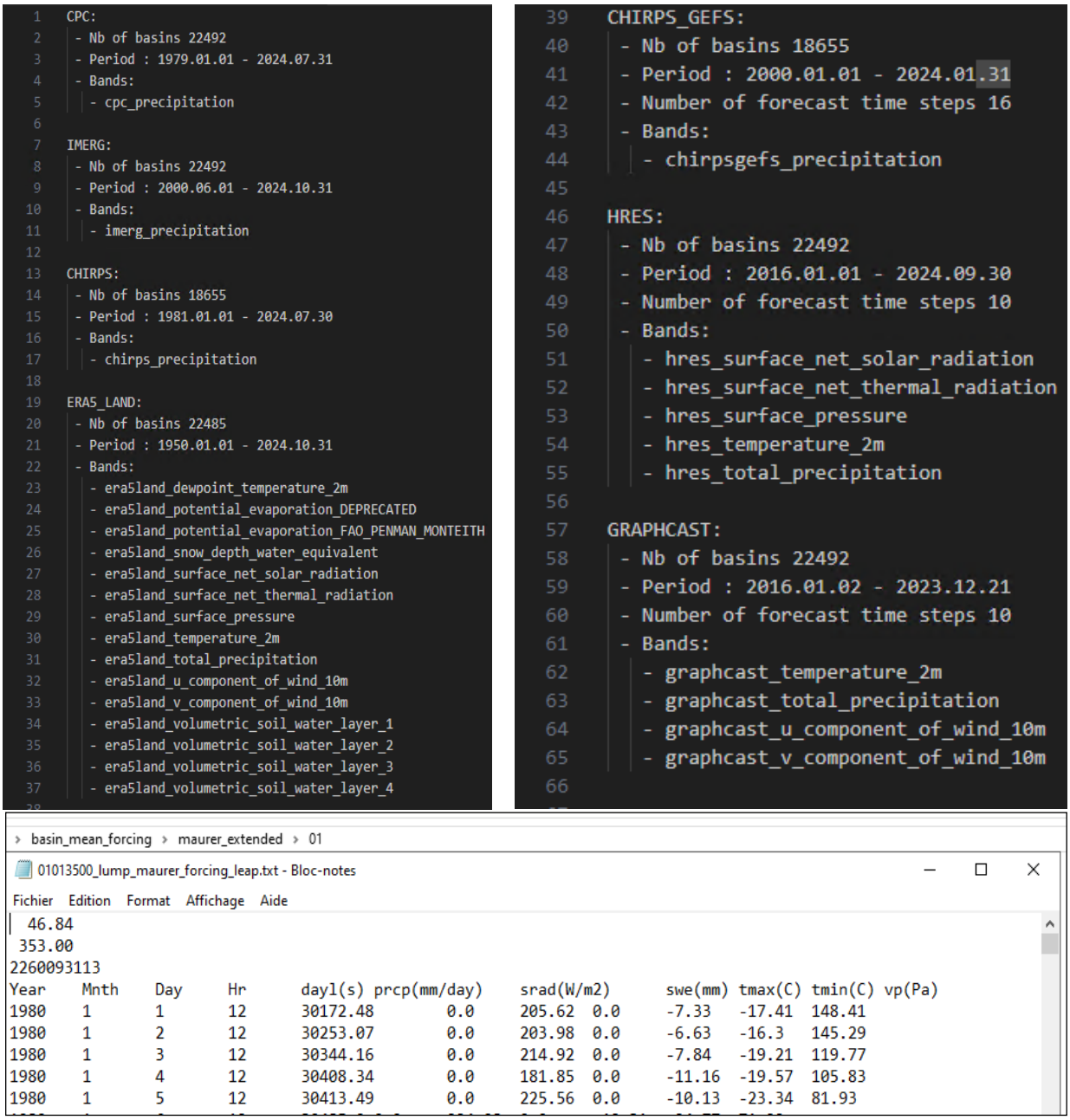
Furthermore, you are using climatological ensembles to create a possible forecast, but I do not believe this is the best strategy to do that. Climatology is used, normally, in medium to seasonal range (so couple of weeks to some months) where the forecast models of the meteorological variables are no longer reliable, and for variables that present a cyclic pattern (temperature, radiation…) but you are using it in precipitation 1 or 3 days ahead, which I do not believe is a good practice. How is the precipitation of the 1st of November for the last 18 years related to the 1st of November of this year? I do not think there is a strong relationship between these values that can be used for short-term forecasting, especially for precipitation, which is the most important variable to drive the forecast. If you want to use it for temperature or radiation, that can be an (non-ideal but defendable) option, but I would highly recommend to not use it for precipitation.

The use of climatological ensemble as a no-skill benchmark for meteorological forecasts is a common practice to evaluate hydrometeorological forecast performances. Figure 4 illustrates that the climatological ensembles distribution does not reveal any evident bias for both the rainfall and the PET despite the simplicity of their construction.

In this section, you also indicate the absence of operational weather forecast archives, but this is also not fully correct. Shalev (2024) release a historical weather forecast for CARAVAN, which includes the CAMELS US dataset. It is true that not all the products are available in the testing periods that you have but: One option is to use CHIRPS-GEFS which is only precipitation but is available in your period of interest. Another, better option, is to benchmark a simple LSTM against Kratzert 2019, and once that is working well, you evaluate your model and the new LSTM in the periods where you have a historical forecast (2016-2024) from Shalev. This is more work but would actually give you a robust study to evaluate your model under real forecast conditions.

We thank the reviewer for mentioning this study and the associated datasets. Nevertheless, the use of the meteorological forecasts (and nowcast) products provided by Shalev and Kratzert (2024) in the present study presents major challenges, several of which were identified by the reviewer. First, CHIRPS-GEFS provides only precipitation forecasts and covers the training but not the testing period that should be used in the revised version of the manuscript. Its integration would require generating realistic forecast scenarios for the remaining forcing variables, applying an appropriate bias-correction procedure to precipitation and rerunning the entire training/testing workflow process for all tested models. In addition, no overlap period exists between the meteorological forcing series used in the present study (Maurer extended version) and the historical forecasts products available on the 2016-2024 period. **Figure A** summarizes the availability of the

variables between the products suggested by the reviewer and the datasets used in the present work.



```
CPC:
 - Nb of basins 22492
 - Period : 1979.01.01 - 2024.07.31
 - Bands:
   - cpc_precipitation

IMERG:
 - Nb of basins 22492
 - Period : 2000.06.01 - 2024.10.31
 - Bands:
   - imerg_precipitation

CHIRPS:
 - Nb of basins 18655
 - Period : 1981.01.01 - 2024.07.30
 - Bands:
   - chirps_precipitation

ERA5_LAND:
 - Nb of basins 22485
 - Period : 1950.01.01 - 2024.10.31
 - Bands:
   - era5land_dewpoint_temperature_2m
   - era5land_potential_evaporation_DEPRECATED
   - era5land_potential_evaporation_FAO_PENMAN_MONTEITH
   - era5land_snow_depth_water_equivalent
   - era5land_surface_net_solar_radiation
   - era5land_surface_net_thermal_radiation
   - era5land_surface_pressure
   - era5land_temperature_2m
   - era5land_total_precipitation
   - era5land_u_component_of_wind_10m
   - era5land_v_component_of_wind_10m
   - era5land_volumetric_soil_water_layer_1
   - era5land_volumetric_soil_water_layer_2
   - era5land_volumetric_soil_water_layer_3
   - era5land_volumetric_soil_water_layer_4
```

```
CHIRPS_GEFS:
 - Nb of basins 18655
 - Period : 2000.01.01 - 2024.01.31
 - Number of forecast time steps 16
 - Bands:
   - chirpsgefs_precipitation

HRES:
 - Nb of basins 22492
 - Period : 2016.01.01 - 2024.09.30
 - Number of forecast time steps 10
 - Bands:
   - hres_surface_net_solar_radiation
   - hres_surface_net_thermal_radiation
   - hres_surface_pressure
   - hres_temperature_2m
   - hres_total_precipitation

GRAPHCAST:
 - Nb of basins 22492
 - Period : 2016.01.02 - 2023.12.21
 - Number of forecast time steps 10
 - Bands:
   - graphcast_temperature_2m
   - graphcast_total_precipitation
   - graphcast_u_component_of_wind_10m
   - graphcast_v_component_of_wind_10m
```

basin_mean_forcing › maurer_extended › 01

01013500_lump_maurer_forcing_leap.txt - Bloc-notes

Fichier Edition Format Affichage Aide

```
 46.84
 353.00
2260093113
Year    Mnth    Day     Hr      dayl(s) prcp(mm/day)    srad(W/m2)      swe(mm) tmax(C) tmin(C) vp(Pa)
1980    1       1       12      30172.48        0.0     205.62  0.0     -7.33   -17.41  148.41
1980    1       2       12      30253.07        0.0     203.98  0.0     -6.63   -16.3   145.29
1980    1       3       12      30344.16        0.0     214.92  0.0     -7.84   -19.21  119.77
1980    1       4       12      30408.34        0.0     181.85  0.0     -11.16  -19.57  105.83
1980    1       5       12      30413.49        0.0     225.56  0.0     -10.13  -23.34  81.93
```

Figure A: Availability between the needed archives and those provided in Shalev and Kratzert (2024). Screenshot are shown for the nowcast products (top-left), forecast products (top-right) up to our covered lead times, and an example of the Maurer extended forcing variables (bottom) which stops in 2008.

For these reasons, the use of these meteorological forecast series appears as a research work in itself, beyond the scope of this manuscript. The text of the manuscript will be modified to mention the existence of these historical forecast databases and explain why they were not used in this work.

Section 2.2 and 2.5.2

Here you indicate that you use an ensemble of 60 runs. Why 60? Most studies use between 5 and 10. Do you get a significant different with 60? If you want to use 60 that is your choice, but in section 2.5.2 you indicate that because of the 60 ensembles you have unreasonable computational cost, and that you are only going to use a subset of basins. If 5 to 10 ensembles give you the same as 60, then you can reduce the computational costs, and then do the study in the full region, which would produce more robust result.

Also, the 18 ensembles members can be accommodated in the batch dimension of the tensor and the different seeds can be run in parallel (even in a single GPU) so most of the computational overhead you are reporting can be overcome with some technical tweaking.

We thank the reviewer for this insightful remark. Our initial choice to use a large number of seeds aimed at thoroughly representing the modeling uncertainty. This choice ultimately proved computationally demanding, particularly for the climatological ensemble evaluation. Following the reviewer's recommendation, we will limit the revised experiments to 20 seeds. However, we will keep the same sample of 56 basins for the climatological evaluation, as it remains sufficiently representative of the full set of 531 basins.

Line 38: I do not agree that discharge simulation and discharge forecasting are fundamentally different tasks. You are trying to model the same system and the same rainfall-runoff response. In forecasting mode, you have the increased uncertainty of the meteorological input, however that is more of a limitation and not a fundamentally different task. Multiple operational models are calibrated with observed data in pseudo-forecast mode and later incorporated in forecasting pipelines, and they work well. Data-driven methods give the advantage that, if trained with real forecast, they can learn to compensate for systematic biases, but again, this is more of training strategies to compensate for data quality limitations and not because the task is fundamentally different.

We thank the reviewer for this remark. We will revise this paragraph (lines 38-46) in the final version.

Line 66. The tittle of Fig 1 should be more self-explanatory.

The legend of Figure 1 will be modified: "Cumulative density functions of the Nash-Sutcliffe and persistence criterions obtained for two types of rainfall-runoff models - the conceptual model SACSMA (Newman et al., 2017) and the LSTM proposed by Kratzert et al. (2019). The

two models have been trained and evaluated on the CAMELS-US dataset. The illustrated scores correspond to the independent evaluation period."

Line 127-128: What do you mean by: "The direct forecasts from the benchmark models were assumed to be unchanged for the tested lead time; therefore, no further running was necessary".

To remove any ambiguity, this sentence will be reformulated as follows: "The forecasts of the initial models are used as inputs of the proposed DA approaches ".

Figure 7. You should explain the colors also on the legend of the figure and not only on the text above. The figure plus the legend should be self-explanatory. Also, as a suggestion the message of this figure would be better explain by a boxplot per lead time graph. The boxplot would give the distribution along the basins and because you have one for each model and for each lead time it can be easily compared. Something similar to your Figure 8, but for the different lead times (you can also see Figure 3 from Nearing 2024).

The same color will be used in all graphics in Figure 7. Moreover, Figures 7 and 8 present complementary illustrations of the results. ECDF-plots such as in Figure 7 are common in previous works and highlight differences between models, corresponding to shifts of the ECDF, more clearly than boxplots. The boxplots on the other hand, show to which extend the gains are systematically observed for all watersheds or not.

Section 3.2.2: Can you explain in detail how did you constructed this figure? How did you construct the 10 classes? What does a lower or a higher rank indicates?

The tested (DA) approaches differ in ensemble sizes, with some configurations using 18 x 8 members (forecast members and seeds) and others using 18x60 members. For the ease of comparison, all ensembles were reduced into 10 equiprobable classes for the rank diagrams.

Technically, the ranks of the observed discharge relative to the ensemble forecast were calculated using the *rank_histogram()* function from the **xskillscore** library, which is based on Hamill (2001). The so-obtained rank values were then grouped into 10 bins using the *groupby_bins()* method applied to the corresponding *xarray.DataArray* object, with the **bins** argument set to 10. Each bin was assigned a categorical label (1-10), and the frequency of occurrences in each category was computed over the evaluation period for each basin.

The lowest class represents the frequency of the observed discharge falling lower than the ensemble's 10^{th} expected percentile, whereas the highest class corresponds to observations exceeding the 90^{th} percentile.

To assess the potential impact of differences in ensemble size across approaches, we constructed an illustrative example based on a synthetic ensemble drawn from a Gaussian distribution N(0,1). Using 1000 realizations and 30 independent resampling, we compared the 10-class rank histograms obtained with ensembles of 100 and 2000 members. As shown in **Figure B**, and as expected, both ensemble sizes produce uniform distributed rank diagram.

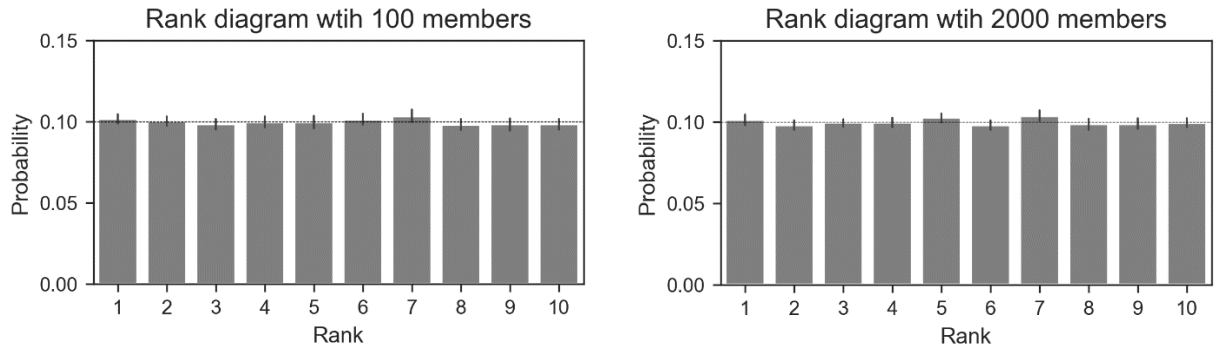


Figure B: Example of a 10-class rank diagram constructed from a Gaussian distribution (N (0,1)) generating an ensemble of 100 members (left) and 2000 members (right), with 1000 examples each. In each case, the last member (100th or 2000th) is used as the verification value. Error bars indicate variability across 30 independent random realizations, while uniformity line is shown with the horizontal dotted black line.

Line 330-335: Can you explain in detail how are you evaluating the LSTM here to produce these results? Also, you are indicating that "This result suggests that the LSTM model is insufficiently responsive to recent meteorological inputs". Can it be that because the LSTM is driven only by meteorology, and because the climatological forecasts are not good (see my comment above about that) then the predictions are biased? The other models have the advantage of having discharge, which is a highly autoregressive variable, so they somehow compensate. However, if all you have is meteorological forecast and these are non-sense, how can the model perform well? I think this point is important and can biased the results you are presenting.

We acknowledge the reviewer's concern; but none of the presented results indicate any significant bias, either in the meteorological ensembles or in the produced discharges forecast.

As illustrated in Figure 4, and further supported by the rank diagrams of the other tested models in Figure 10, the proposed climatological ensembles are widely spread (reflecting no-skill meteorological forecast) but remain neither biased nor of non-sense. Similarly, the rank diagrams (Figure 10) do not indicate any systematic bias in the prediction for any of the tested methods, including the LSTM. A minor bias, a slight overestimation, may be observed in the rank diagram of the MLP Simple approach. The observed differences in the model's skills are

therefore attributable not to forecast bias but to the reliability of the ensembles (i.e., of their spread), which inevitably affects the forecast resolution.

In the context of ensemble forecasting, the LSTM appears penalized when fed with no-skill ensembles due to the relatively low dispersion of its predicted ensemble members. This reduced spread, compared to the other approaches, is clearly visible on Figure 11 as revealed the rank diagrams in Figure 10.

It is also important to note that when meteorological forecasts have little or no skill, the model's performances rely primarily on the hydrological inertia of the watersheds, a factor accounted for in all tested approaches, including the original LSTM.

Finally, the paragraph in lines 330-335, which appears awkwardly written, will be reformulated or eliminated in the revised version.

References

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., & Nevo, S. (2022). Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. Hydrology and Earth System Sciences, 26(21), 5493–5513. https://doi.org/10.5194/hess-26-5493-2022

Nearing, G., Cohen, D., Dube, V. et al. Global prediction of extreme floods in ungauged watersheds. Nature 627, 559–563 (2024). https://doi.org/10.1038/s41586-024-07145-1

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resources Research, 56,e2019WR026793. https://doi.org/10.1029/2019WR026793

Yang, Y., Pan, M., Feng, D., Xiao, M., Dixon, T., Hartman, R., Shen, C., Song, Y., Sengupta, A., Delle Monache, L., & Ralph, F. M. (2025). Improving streamflow simulation through machine learning-powered data integration and its potential for forecasting in the Western U.S. Hydrology and Earth System Sciences, 29(20), 5453–5476. https://doi.org/10.5194/hess-29-5453-2025

Shalev, G., & Kratzert, F. (2024). Caravan Multi Met: Extending Caravan with multiple weather nowcasts and forecasts. arXiv preprint arXiv:2411.09459. https://arxiv.org/abs/2411.09459

Kratzert, F. (2019). CAMELS Extended Maurer Forcing Data, HydroShare, https://doi.org/10.4211/hs.17c896843cf940339c3c3496d0c1c077

Hamill, T. M., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Mon. Wea. Rev., 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings, ECMWF Workshop on Predictability, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.].