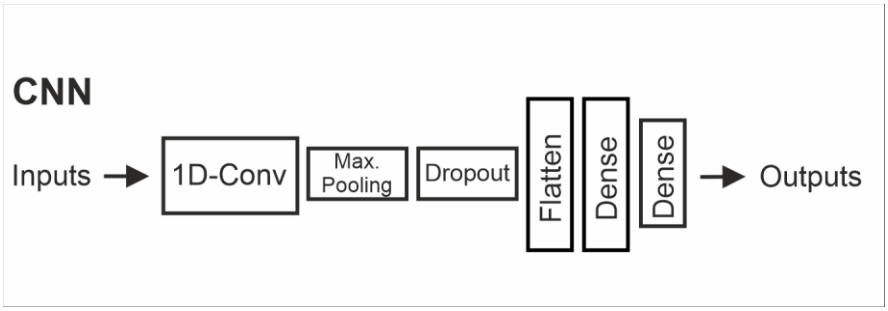
We thank the referee for their careful reading and helpful comments. Our reply is given below.

2.1 In Data and study sites section (2.1 to 2.5) (p3. From 85-165): it could be better to explain the reason behind using different potential evapotranspiration (PTE) calculation methods for different basins and the impact of different methods on PTE, if there is any.

We agree that it is better to explain why different methods are being used for calculating potential evapotranspiration. We think that there is no negative impact of using different methods; on the contrary, the methods are carefully chosen to provide the most relevant estimations, taking into account the available meteorological data, the climate of the area and local expert knowledge. The manuscript was modified accordingly, L261: *"For reservoir models, evapotranspiration processes were considered using time series of potential evapotranspiration, which were calculated for each site using different methods depending on the available meteorological data, the climate of the area and local expert knowledge."*

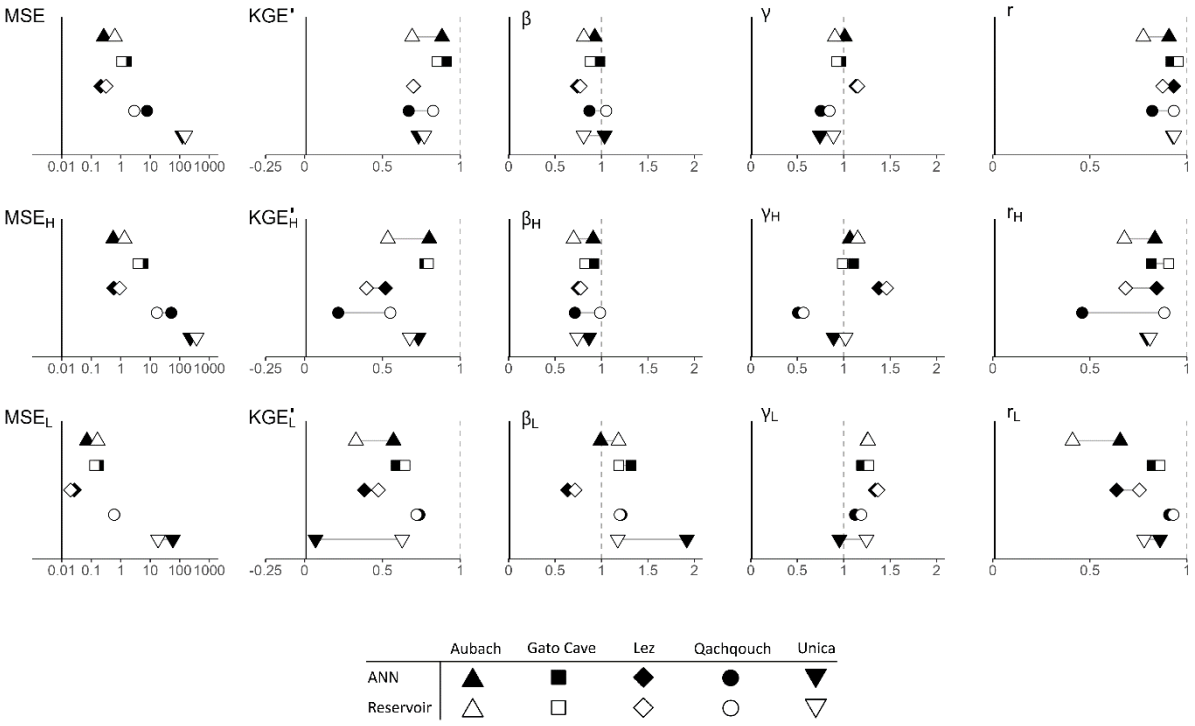
2.2 In Artificial neural networks section (3.1) (p6. From 175 to 185): Selected ANN model type should better be demonstrated with figures as it is done for reservoir models in the following section. Since CNNs are quite complex deep learning models, it can be difficult to be comprehended even for experienced ANN modelers. Demonstrations of CNN and a little more detailed explanation of the regularization methods to avoid traps such as exploding gradient could enrich the context of the article.

A figure has been added to better explain the functioning of ANN models. Further details have also been added to the description of the ANN modelling approach in section 3.1: *"To ensure proper learning, the models are regularised with several measures. Hence, early stopping with a patience of 20 Steps is applied to prevent the model from overfitting. Except Qachqouch, where little data is available, the size of the according stopset ranges between one and four annual cycles (see provided scripts for details). This stopset is considered a part of the calibration period mentioned in Sect. 3.4. Further, dropout ensures robust training and serves as another measure against overfitting. We applied the Adam optimizer for a maximum of 150 to 300 training epochs with an initial learning rate of 0.001 and applied gradient clipping to prevent exploding gradients."*



2.3 In Model evaluation section section (3.4) Model evaluation criteria should be better justified. Moreover, using more evaluation criteria can be more rigorous which are usually considered in hydrological modelling studies. For instance, relative volume bias, normalized peak error, root mean square error, nash score (the last two is suggested sice they are also used as cost function).

Thanks for the recommendation, we added the MSE to the list of evaluation criteria as it is the performance metric used for calibrating the ANN and reservoir models. Now the models are evaluated using MSE, modified KGE, Diagnostic Efficiency, β for the volume bias, γ for the discharge variability and the Pearson correlation coefficient r for the discharge shape and timing – Figure 4 was updated (see below). We also explained in more details why we evaluate the results using the modified KGE instead of the Nash-Sutcliffe Efficiency: *"The Kling-Gupta Efficiency (KGE) has gained in popularity as it aims to address some limitations of the Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970), i.e. (i) the discharge variability is underestimated, (ii) the mean of observed values is not a meaningful benchmark for variables with high variability, and (iii) the normalised bias is dependent to the variability (Gupta et al., 2009, Willmott et al., 2012)."*



2.4 In Introduction section (p. 2) from line 48 to 51, authors said "Distributed models require a lot of data for defining physical parameters and thus can 50 be tough to use in a scarce data context. On the other hand, data-driver models permit studying complex and heterogeneous karst systems without requiring extensive meteorological and system-related data.". However, the difference between distributed and data-driven models (if the authors meant statistical models such as ANN) about the data requirement is not necessarily about the amount of data but rather about the diversity. This phrase can be modified to indicate that distributed models need more diverse data while ANNs need only input and output data.

We thank the reviewer for the suggestion. The sentence has been modified to emphasise that distributed models need more diverse data (both with high spatial and temporal resolution), in contrast to lumped parameter models:

L49: *"Distributed models require a lot of diverse data with high spatial and temporal resolution for defining physical parameters and thus can be tough to use in a scarce data context (Hartmann et al., 2014)."*

L50: *"On the other hand, data-driven models permit studying complex and heterogeneous karst systems without requiring extensive meteorological and system-related data with high spatial resolution."*

2.5 In Introduction section (1) (p. 2) from line 57 to 62 references can be given in historical order. In Introduction section (1) (p. 3) in line 74 references can be given in historical order. In Reservoir model section (3.2) (p. 7) in line 74 references can be given in historical order. [References can be homogenized as so...]

All the references were ordered in historical order.

2.6 In Source of uncertainties section (4.2) (p.23) from line 518 to 540, authors can mention the expected (if not quantified) amount of uncertainty for each source in their case studies for different basins either by expertise or literature support. Since basins are of different characteristics, it can be interesting to have such information for readers.

Such information is unfortunately not available for the basins of this study, either by expertise or literature support. However, in the literature, several authors focused their work on quantifying the uncertainties related to input and observed data in the context of karst aquifers. We have added details about this for each source of uncertainty in section 4.2:

Input data: *"McMillan et al. (2018) suggested that uncertainties in precipitation data are about 0–10 % at point scale but can go up to 40 % when considering interpolation uncertainties."*

Observed data: *"The uncertainties related to discharge measurements are highly dependent on the quality of the gauging station and usually range between 10–40 % (McMillan et al., 2018). Although they are expected to be higher in a karst context (Westerberg et al., 2016), some authors reported uncertainties of about 20 % (Jeannin et al., 2021) or 10–15 % (Katz et al., 2009)."*

To the best of our knowledge, quantitative information about model structure uncertainties in karst environments has never been explicitly mentioned.

References

Jeannin, P.-Y., Artigue, G., Butscher, C., Chang, Y., Charlier, J.-B., Duran, L., Gill, L., Hartmann, A., Johannet, A., Jourde, H., Kavousi, A., Liesch, T., Liu, Y., Lüthi, M., Malard, A., Mazzilli, N., Pardo-Igúzquiza, E., Thiéry, D., Reimann, T., Schuler, P., Wöhling, T., and Wunsch, A.: Karst modelling challenge 1: Results of hydrological modelling, J. Hydrol., 600, 126508, https://doi.org/10.1016/j.jhydrol.2021.126508, 2021.

Katz, B. G., Sepulveda, A. A., and Verdi, R. J.: Estimating Nitrogen Loading to Ground Water and Assessing Vulnerability to Nitrate Contamination in a Large Karstic Springs Basin, Florida1, JAWRA Journal of the American Water Resources Association, 45, 607–627, https://doi.org/10.1111/j.1752-1688.2009.00309.x, 2009.

McMillan, H. K., Westerberg, I. K., and Krueger, T.: Hydrological data uncertainty and its implications, WIREs Water, 5, e1319, https://doi.org/10.1002/wat2.1319, 2018.

Nash, J. E. and Sutcliffe, J.: River flow forecasting through conceptual models: Part 1. A discussion of principles., J. Hydrol., 10, 282–290, 1970.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., and Freer, J.: Uncertainty in hydrological signatures for gauged and ungauged catchments, Water Resources Research, 52, 1847–1865, https://doi.org/10.1002/2015WR017635, 2016.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, Intern. J. Climatol., 32, 2088–2094, https://doi.org/10.1002/joc.2419, 2012.