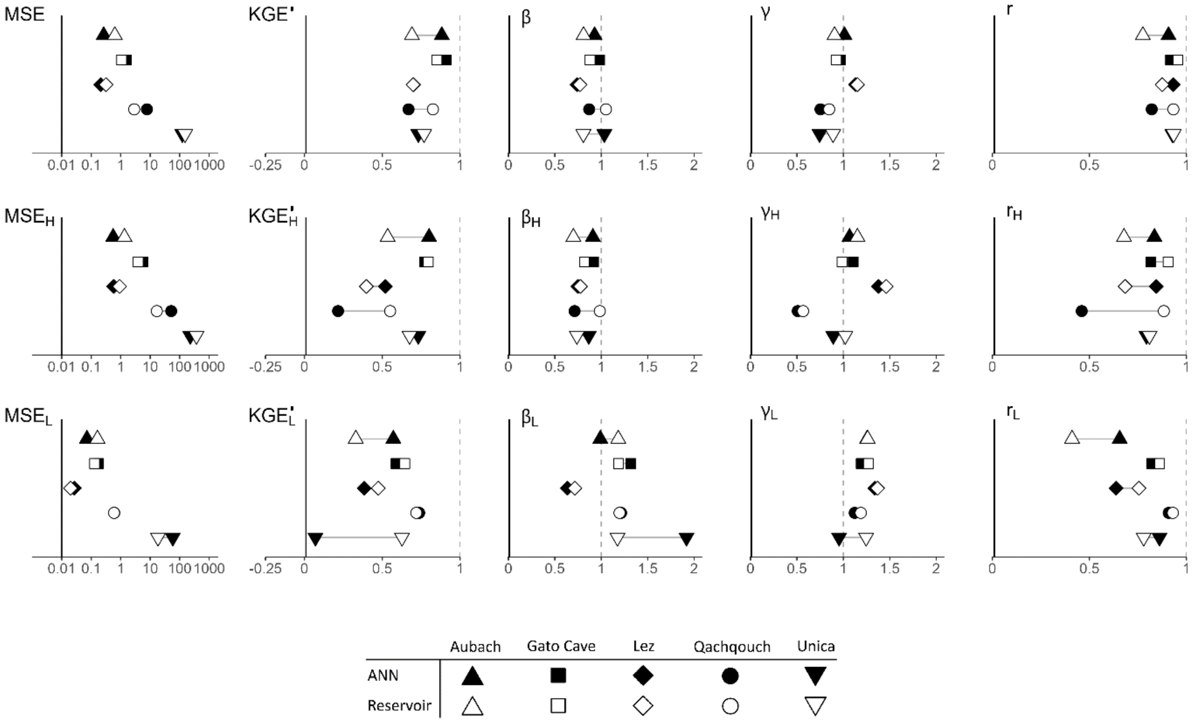
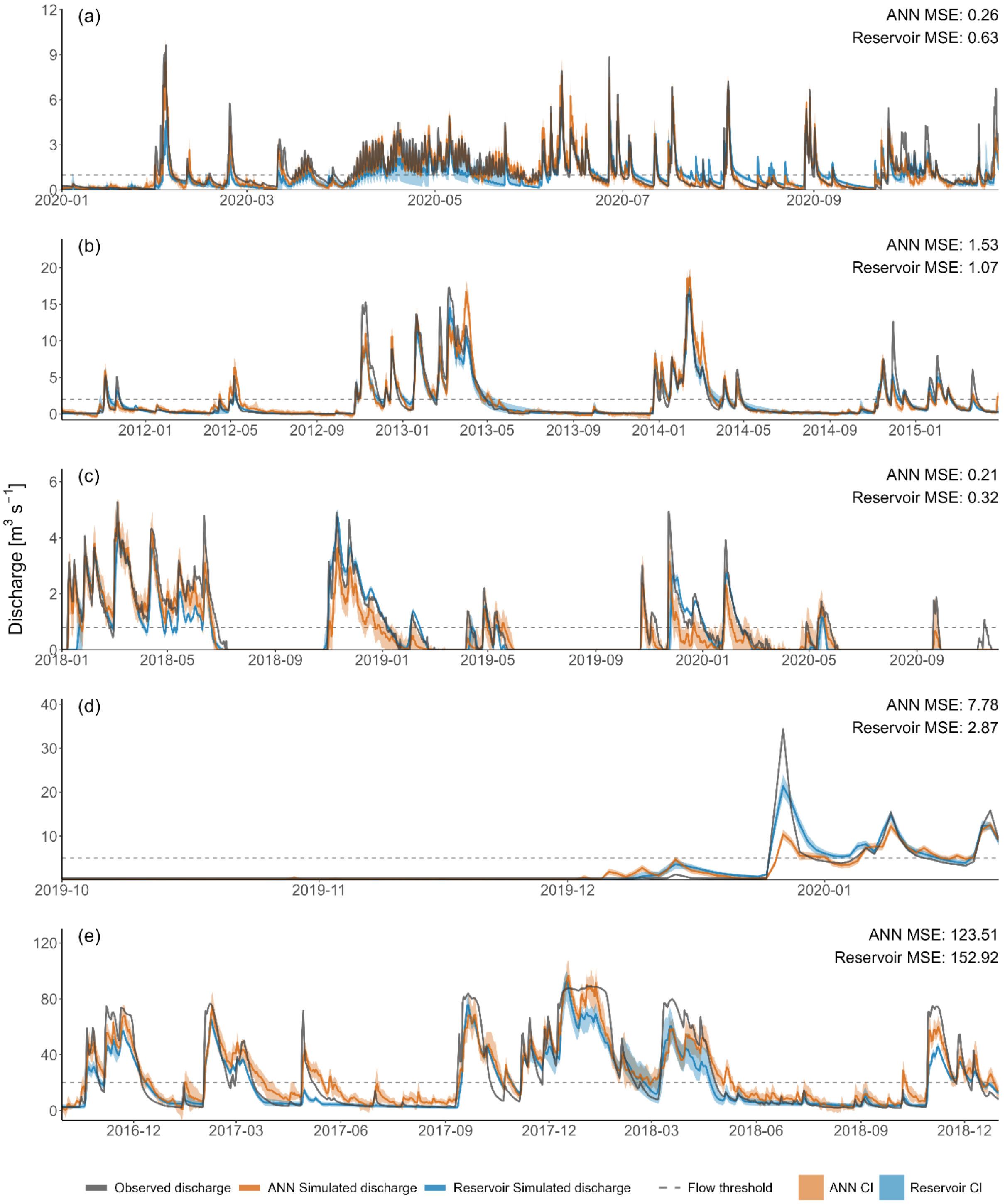
We thank the reviewer for the very relevant major comments, which helped us to improve the quality and relevance of the models. To address comment 1.1 (use same performance metric for the calibration) and comment 1.5 (calibration of the parameters of the snow routine), we reran the reversoir models using KarstMod in command line, thus allowing (i) to use the same performance metric (MSE) for both modelling approaches and (ii) to calibrate the parameters of the snow routine.

1.1 Line 290: I think the authors must explain why they used different performance metrics for the calibration of the models. Choosing the objective function(s) is the subjective decision of the modeller, however, if you want to make a comparison between the model performances, I recommend calibrating the models according to the same performance metric.

Initially, we used different performance criteria because MSE is the most common metric for calibrating ANN models but is not implemented in KarstMod. However, we agree with the reviewer that the comparison of the results will be more relevant using the same performance metric for calibration. We used KarstMod in command line, thus allowing to use MSE as a performance metric for the reservoir models. Both ANN and reservoir models are now calibrated using the same performance metric. The figures 2, 3, 4, 5, the table 4 and the text of section 4 have been updated with the new results and also include the MSE for the evaluation of model performance.




(a)
ANN MSE: 0.26
Reservoir MSE: 0.63
0
3
6
9
12
2020-01
2020-03
2020-05
2020-07
2020-09
(b)
ANN MSE: 1.53
Reservoir MSE: 1.07
0
5
10
15
20
2012-01
2012-05
2012-09
2013-01
2013-05
2013-09
2014-01
2014-05
2014-09
2015-01
(c)
ANN MSE: 0.21
Reservoir MSE: 0.32
Discharge [m^3 s^{-1}]
0
2
4
6
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
2020-01
2020-05
2020-09
(d)
ANN MSE: 7.78
Reservoir MSE: 2.87
10
20
30
40
2019-10
2019-11
2019-12
2020-01
(e)
ANN MSE: 123.51
Reservoir MSE: 152.92
0
40
80
120
2016-12
2017-03
2017-06
2017-09
2017-12
2018-03
2018-06
2018-09
2018-12
Observed discharge
ANN Simulated discharge
Reservoir Simulated discharge
Flow threshold
ANN CI
Reservoir CI

1.2 Line 294: Is this adequate for a successful calibration?

Indeed, this can be improved for the reservoir modelling approach. As suggested in comment 1.3, we reran the reservoir models during a 6-hour period and selected all the simulations above a specific threshold value for the objective function (MSE).

1.3 Line 298: This is not a scientific statement. Running the model for 1 hour does not tell anything about the number of simulations. Instead, I recommend applying a threshold value (e.g. NSE > 0.5) for the behavioural runs and selecting all of the simulation results above this threshold rather than selecting the top 1000 simulations. By doing this, you would probably get a different number of behavioural model runs in each catchment and you can better compare the model uncertainty bounds. (Alternatively, you may mention the stop criteria of the KarstMod model e.g. simulation-time limitation, or number of behavioural runs)

Thank you for the relevant comment. We indeed used a threshold value for selecting all the simulation results but this was poorly phrased in the manuscript. We reran all the models using a maximum MSE threshold on the calibration period for a 6-hour model run and modified the manuscript accordingly, L298: *"In KarstMod (reservoir models), the retained simulations correspond to all the results satisfying a maximum MSE threshold on the calibration period for a 6-hour model run."*

1.4 Line 305: Here you mention that you evaluated the model performances by using KGE and DE, but in line 290, you calibrated the models by using NSE (reservoir model) and MSE (ANN model) metrics. So, why did you change the objective function in the model evaluation phase? Please explain. As far as I experienced, the best simulation obtained by any of the performance metrics (e.g. NSE) does not guarantee the best simulation on another metric (eg. KGE). So, please calibrate and evaluate the models again by using the same performance metric(s).

Thank you for the comment. As explained in comment 1.1, we calibrated and evaluated the models again using the same performance metric (MSE). We also added the MSE results for the validation period in Figure 4 and Table 4. However, we keep the in-depth evaluation of the models results with the modified KGE and its components, as it provides valuable insights into the different aspects of a model (i.e. variability, volume and correlation).

1.5 Line 350: This plot shows that the reservoir model outperforms the ANN model in Spanish, French, Lebanese and Slovenian catchments except for the Austrian catchment. I think the Austrian catchment is dominated by snow (as you mentioned in line 379) and the reservoir model structure is not adequate to beat the ANN model. It must be discussed in the discussion part. Additionally, your snow routine may not be adequate if you did not calibrate the snow parameters. Please see the paper (Çallı et.al. 2022).

Please see comment 1.18 about the model structure of the Aubach model. Following your suggestion (also comments 1.19 and 1.24), we reran the reservoir models including the estimation of the parameters of the snow routine by model calibration. Despite the catchments (except Aubach) not being hugely affected by snow, this helped to get more relevant results while strengthening the methodology of the reservoir modelling approach.

1.6 Lines 39-40: In my opinion, you would have a better classification of the models. Kovacs and Sauter (2007) do not classify the models as "data-driven" or "distributed" models. Distributed or lumped, all models are somehow data-dependent. I think that would be better to classify the models as "black-box models", "conceptual models" and "physical models" considering their complexities. You may give details about the "Machine learning models" under the "Black Box", and reservoir models (or lumped parameter models) under the "Conceptual models". You may also give some details about the advantages and disadvantages of these modelling approaches regarding the complexities and data requirements. You may mention why so many researchers choose conceptual models. Another point is that, please be more consistent about the model classification inside the paper.

We agree, the classification of modelling approaches as "distributed" or "lumped parameter" is better suited and consistent with the classification of Kovacs and Sauter (2007). We modified the manuscript accordingly, also giving details about (i) the neural networks models under the "black box" class, and (ii) the reservoir models under the "lumped parameter" class, L43: *"They include (i) "black-box" models such as neural networks-based approaches, which use no a priori information about the functioning of a system; and (ii) "conceptual" models, which are based on a conceptual representation of a karst system – e.g., for the reservoir models, a succession of one or several reservoirs using simplified physical transfer functions."* The advantages and disadvantages of these modelling approaches are extensively detailed and discussed in section 4.3 (Comparison of general model properties). We added a sentence in the introduction to explain why conceptual models such as reservoir models are well suited to the study of karst systems: *"This approach is well suited to karst systems due to the high heterogeneity and low level of knowledge of their structure (Fleury et al., 2009; Hartmann et al., 2012)."*

1.7 Line 47: You may consider citing the paper (Addor and Melsen, 2019) about the model selection procedure (adequacy or legacy).

Indeed, this is an important factor of choice. We modified the manuscript accordingly, L47: *"The choice of a modelling approach depends mainly on the objective of the study, but also on the current knowledge of the system, the available data and regional/institutional preferences (Addor and Melsen, 2019)."*

1.8 Line 50: You may consider rephrasing the sentence "…distributed models require a lot of data". You may alternatively say: "Distributed models require the data with high spatial resolution, however, lumped models require data in high temporal resolution." (You may cite here again Hartmann et.al. 2014 or Kovacs and Sauter 2007).

We thank the reviewer for the suggestion. Indeed, it is relevant to detail what type of data is needed in distributed and data-driven models. We modified the manuscript accordingly.

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."*

1.9 Line 51: I would remove the sentence "Both black-box and reservoir models …." to avoid repetition. The following sentences already explain the applications of the ANN and reservoir models for academic and operational purposes.

As suggested, the sentence was removed.

1.10 Lines 55-60: Please be more consistent about the references in the brackets. You may use time order (Perrin et.al. 2003; Jukic and Denic-Jukic 2009; Tritz et.al 2011; Bittner et.al. 2020) or alphabetical order (Bittner et.al. 2020; Jukic and Denic-Jukic 2009; Perrin et.al. 2003; Tritz et.al 2011).

All the references in brackets were ordered by time order.

1.11 Line 81: The paper does not include any simulations by using the artificial future data (different emission scenarios) to compare the model adaptability against climate change. So this is not fair to have such an inference. If you want to declare that the models are not robust in extreme event predictions, please cite several climate-related studies (to better link the connection between the climate-change and extreme events).

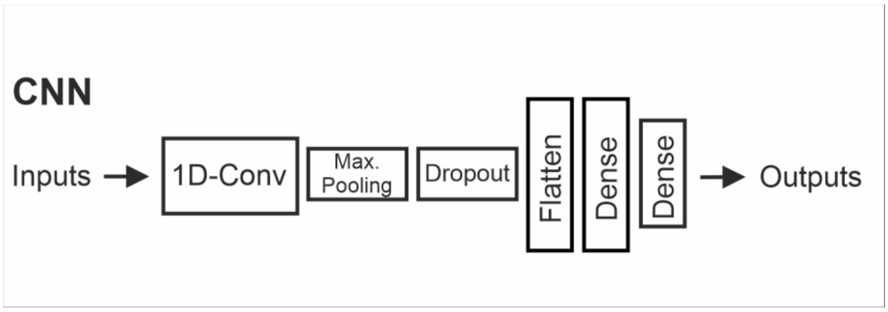
We agree that it is not appropriate to provide inferences on the models' suitability for climate change predictions. We removed the sentences concerned (in the introduction and conclusion).

1.12 Line 91: I recommend moving map A.1 to this section.

We moved Figure A1 to section 2 (main text, now Figure 1).

1.13 Line 174-175: In this sub-section, adding a schematic illustration of the ANN model would be very helpful to better understand the modelling architecture.

A figure was added to explain the modelling architecture of the ANN model.



1.14 Line 180: Please simply explain why you applied the 1-D convolutional layer approach (you may consider citing previous ANN modelling studies).

A sentence was added in the introduction to explain why we apply 1D Convolutional Neural Network: *"We specifically apply 1D Convolutional Neural Networks (CNNs) because in an earlier study (Wunsch et al., 2022) we were able to demonstrate their high ability to perform karst spring discharge modelling. Furthermore, they have some favourable properties compared to popular recurrent neural networks (e.g. the LSTMs), such as batch-wise training procedure which makes them considerably faster and computationally less expensive."*

1.15 Line 184: I think you can move the names of the python libraries to the appendix. It is not necessary inside the text.

The names of the Python libraries were moved in the *Acknowledgements* section.

1.16 Line 185: Here I see you used the library BayesOpt. Was that library used for the Bayesian uncertainty analysis? Or did you apply an uncertainty analysis to the ANN model predictions? If so, please explain which method you applied. Please give some details about the model assumptions, parameter distributions etc.

Actually, we used the library Bayesian Optimization, not BayesOpt. We apologise for this inaccuracy. We did not perform a Bayesian uncertainty analysis with this package, but used it to optimise the ANN hyperparameters. This optimisation strategy unfortunately was not yet explained in the paper. We therefore added clarifying statements and explanations and thank the reviewer for noticing. Please note that the uncertainty is estimated from the model ensemble with 1000 members, which in the case of the CNN are generated by a combination of different model initializations and runs based on monte carlo dropout. A paragraph was added: *"Besides number of filters and number of neurons in the first dense layers, we optimised the training batch size and the length of the input sequence for each simulation step using the Bayesian Optimization library (Nogueira, 2014). The number of minimum and maximum optimisation steps was individually selected for each site and can be found in the provided modelling scripts (Cinkus and Wunsch, 2022)."* The reference of the Python library was corrected, L183: *"[...] Bayesian Optimization (Nogueira, 2014) [...]"*

1.17 Line 189: You may consider pointing out the functionality of the conceptual models in karst water predictions. What is the main advantage of this modelling approach?

Please see the response of comment 1.6. The main advantage of the conceptual models regarding karst water predictions is now detailed in the introduction.

1.18 Line 240: You may consider adding a snow reservoir above the Epikarst in Fig 1a. This would be more suitable for the mountainous catchment.

Indeed, it would be more appropriate. However, this option is not available in the KarstMod platform, and appears to be one of the limitations of using a platform for reservoir modelling (similar to the consideration of polje and surface water influence in the Unica model). We added some details about this aspect both in the Aubach and Discussion sections:

L378: *"These errors can be either due to (i) a miscalibration of the snow routine, retaining too much water as snow in winter and thus releasing too much in warmer periods, or (ii) the snow dynamics which cannot be taken into account within the KarstMod platform, e.g. by adding a snow storage above the epikarst (Chen et al., 2018)."*

L588: *"[...] which may benefit the Qachqouch model; (iii) considering polje and surface water influence in the Unica model; or (iv) considering snow dynamics in the structure of the Aubach model."*

1.19 Line 261: Please mention how to determine the snow routine parameters (Degree-day factor and melting temperature). You mentioned that you did not make an optimization for the snow parameters, so please cite the relevant literature (e.g. He et.al. 2014).

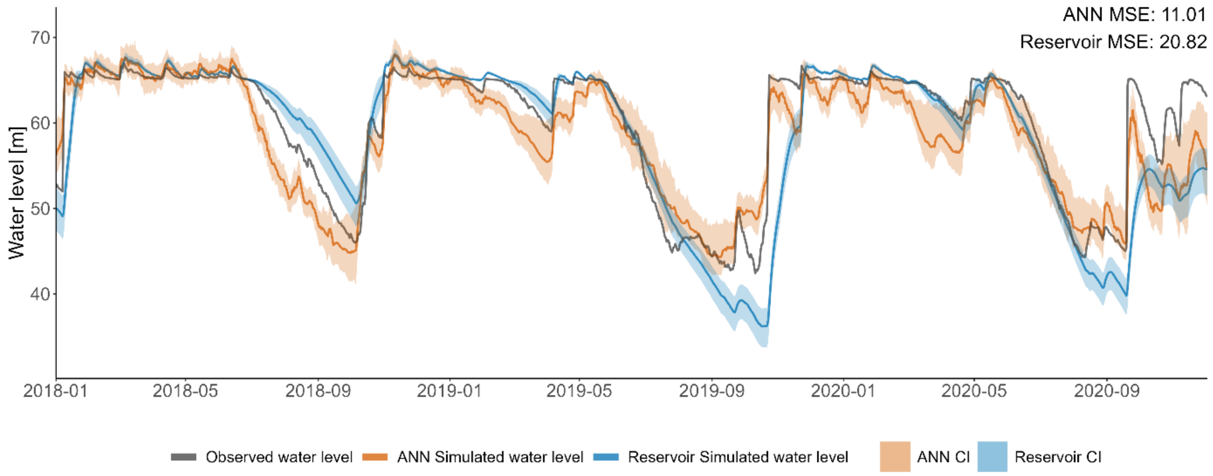
The parameters of the snow routine are now estimated by model calibration. Please see the response to comment 1.5 for more details.

1.20 Line 343: The uncertainty bounds are not easy to see especially for the reservoir model in (a). I think when you select all the behavioural simulations (above the threshold), the uncertainty bound will be much more visible. Then the reader can make a visual comparison between them.

This has been improved with the rerun of the models. Please see the responses of comments 1.1 and 1.3 for more details.

1.21 Line 347: Again the same problem. Please apply a threshold for the reservoir model, and use all the behavioural runs to obtain the uncertainty bounds.

This has been improved with the rerun of the models. Please see the responses of comments 1.1 and 1.3 for more details.



1.22 Line 355: Please share the model calibration skills in the table.

You are right that the information is interesting to provide for reservoir models. However, the score of an ANN model during the calibration period is not relevant as it corresponds to its learning period. Thus, it seems more appropriate to add this detail in appendix to avoid potential confusion for the reader. We added a table that shows the calibration and validation scores of the reservoir models (with a comparison to the ANN models for the validation period).

1.23 Line 355: You mentioned that the ANN models require long time series to learn the functionality of the karst system. On the other hand, reservoir models could be calibrated for relatively shorter periods. But, there are some results to be discussed in detail as below:

Lez and Qachqouch springs simulation results support the hypothesis, but the Aubach catchment does not. We expect better calibration skills in the reservoir model in a short calibration period. However, the ANN model outperforms the reservoir model in Aubach. How do you explain this?

While Aubach and Qachqouch springs have similar lengths for the calibration period (nearly six years, and about four years, respectively), Lez spring has a longer calibration period of slightly more than nine years, which we consider not exactly a "short" period. For Qachqouch, it is explained Line 511 that "even when data are available, there is a significant amount of time without (relevant) discharge, for which no input-output relation can be learned." In comparison, Aubach is a very reactive system with a lot of "relevant" discharge, which benefit the training of the ANN model. The difference between the springs' regime of Aubach and Qachqouch can be appreciated in Figure 2. We modified the manuscript to emphasise the aspect of relevant discharge in short time series:

L473: *"This highlights the strength of conceptual modelling to take into account recharge processes and reservoir replenishment, even on a short dataset with long dry periods."*

L575: *"In contrast, a very short time series or a short time series with long dry periods can be detrimental for the learning of ANN model, which seems to benefit from medium/long periods of relevant discharge (at least 5 years)."*

L25: *"[...] (ii) reservoir models can provide good results even with few years of relevant discharge in the calibration period, [...]"*

1.24 Line 389: You can make a representative snow routing even if you do not calibrate the snow parameters. Please cite the relevant literature.

The parameters of the snow routine are now estimated by model calibration. Please see the response to comment 1.5 for more details.

1.25 Line 527: You may discuss the uncertainty in the temperature data. Temp data strongly affect the timing of the recharge, especially in snow-covered areas (Aubach case).

A sentence was added to discuss the uncertainties related to temperature data in snow-covered areas, L524: *"In the case of snow-covered areas, it can result in strong uncertainties on the timing of snow accumulation and melting (Zhang et al., 2016), and therefore the recharge of the aquifer."*

1.26 Line 547: You may add some other model optimization techniques (e.g. cross-validation, see Wilks 2011).

We agree with the suggestion and modified the manuscript accordingly, L546: *"[...] or to use different model optimisation techniques, such as cross-validation (Wilks, 2011)."*

1.27 Line 559: Please discuss the model's structural adequacy here.

Several sentences were added to discuss the model's structural adequacy:

L556: *"For high-flow periods, results slightly favour the ANN approach (except for Qachqouch spring), with consistently accurate volumes and shape and timing (Fig. 6). ANN models also tend to achieve higher flows than reservoir models (Fig. 4); due to the noticeable/strong karstification of the studied systems, the high occurrence of high discharge data may benefit the learning of the ANN models. On the other hand, reservoir models are more dependent on the relevance and the quality of the input data preprocessing, thus can be more affected by the uncertainties presented in Sect. 4.2, especially regarding high flows."*

L558: *"For low-flow periods, results slightly favour the reservoir approach (except for Aubach spring), with very good estimation of volumes and only a slight overestimation of the hydrological variability (Fig. 6). The conceptual representation of the aquifer with reservoirs and transfer functions may help to simulate the recharge process (especially for inertial systems): a precipitation event will not directly result in a discharge increase at the spring if the reservoir is not fully saturated. On the other hand, ANN models seem to not always account for the time needed for the aquifer to replenish, inducing wave-like behaviours during medium- and low-flow periods (Fig. D1), which can hinder the simulation of low flows."*

References

Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, Water Resources Research, 55, 378–390, https://doi.org/10.1029/2018WR022958, 2019.

Wilks, D. S.: Statistical Forecasting, in: International Geophysics, vol. 100, Elsevier, 215–300, https://doi.org/10.1016/B978-0-12-385022-5.00007-5, 2011.

Zhang, J. L., Li, Y. P., Huang, G. H., Wang, C. X., and Cheng, G. H.: Evaluation of Uncertainties in Input Data and Parameters of a Hydrological Model Using a Bayesian Framework: A Case Study of a Snowmelt–Precipitation-Driven Watershed, Journal of Hydrometeorology, 17, 2333–2350, https://doi.org/10.1175/JHM-D-15-0236.1, 2016.