Reply letter to RC1

Authors: Edouard Patault, Valentin Landemaine, Jérôme Ledun, Arnaud Soulignac, Matthieu Fournier, Jean-François Ouvry, Olivier Cerdan, Benoit Laignel

Manuscript n°: HESS-2020-363

Please find in "black" the reviewer's comments, in "blue" our replies, in 'green' the modification in the revised manuscript.

1. General comments and suggestions

I read the paper with interest. Its English and presentation are not flawless, but any shortcomings do not prevent a good understanding of the contents. The combination of the conceptual (or Expert-Based) WaterSed and the data-driven model makes sense as a way to model the impact of land use/management scenarios on the sediment discharges generated and conveyed in a karstic catchment. The application of the models is well described and appears to have been carefully undertaken. The validation was taken seriously by the authors, even extending part of their analyses to extreme events in order to verify the applicability of the model over a wide range of conditions. Below, I list some of my major concerns. In the attached file the authors can find a few suggestions that may improve the paper, as well as specific questions regarding the methodology and results.

Response: We thank reviewer 1 for the positive feedback, and the constructive review which clearly improve the quality of the manuscript. Here, we addressed a reply to each major concern and/or specific comment.

2. Major concerns

The first major concern is the significance of the paper (which I am not able to judge, being from a slightly different field). Simplistically, the authors just test different scenarios with the WaterSed model (representing the processes until the sinkholes) with a deep neural network that models the effect of the karst network, transforming in sediment inflow to the sinkholes (the result from WaterSed) into sediment that reaches the outlet of the catchment (in this case a water treatment plant). The approach is valid and ingenious, but perhaps not a large breakthrough.

Response: Noted with thanks. We agree that the approach might not be a large breakthrough, consisting in the cascade modelling of two existing modelling tools. However, to the best of our knowledge, this approach has never been proposed in the literature. It is interesting because physically based models are generally used to study sub-surface hydro-sedimentary transfers in karstic regions. But these models require 3-D information on the physical characteristics of the studied karst system which are often not available and limit the model's applicability. This approach allows to assess the impact of multiple land use scenarios on the sediment discharge observed at a water treatment plant. We are confident that our article will have a positive impact for all researchers, drinking water suppliers, and land use planners over the entire North-European loess belt which is impacted by the same erosion processes in a similar environment. They will in turn be able to use this tool to make their own assessments. It would also be useful to let other teams, in different geomorphological contexts, take note of that approach and adapt her for their own studies. One other main contribution of this paper is that the proposed tool is integrated into a decision support system for land use planning. As mentioned by Sit et al. (2020) in their review, is that few DL applications in the water literature included decision-making components. This lack of attention to DL in service of decision-making in water academia presents an opportunity.

Another concern is the need for a Deep Neural Network to model the effect of the karst on the suspended sediments. In a superficial analysis of the results, it seems that for all scenarios but eco-engineering, the sediment discharge at the outlet seems to be almost linearly related to the predictions of the WaterSed model. It would be good to try the same approach with a much more simple multi-linear model. It would be good if the authors could include a scatter plot of the sediment discharge vs runoff series predicted by WaterSed (Fig. 3) and another, even more relevant, of sediment discharge predicted by WaterSed vs sediment discharge observed at the outlet (the water treatment plant).

Response: Noted with thanks. Your concerns are legitimate but before using a DNN in this study, we tried a simpler approach with a multi-linear regression model. The results of the simulation with the regression model are presented (Fig.S1a). We observed that the model was unable to correctly simulate sediment discharge, we observed a high dispersion of the values specifically for the most important events. The results were not surprising considering that the karstic system is complex, and the link between the sediment input simulated at the connected sinkholes by the WaterSed model cannot be linked to the sediment discharge at the spring with a simple linear response (Fig.S2b). The response is non-linear considering multiple processes like fast infiltration, matrix infiltration, underground storage, purge, etc. So, we benchmarked a multi-linear regression model vs the DL model with the same inputs. It appears that the DNN outperformed the multi-linear regression model because the DNN was able to consider the non-linearity of the system.

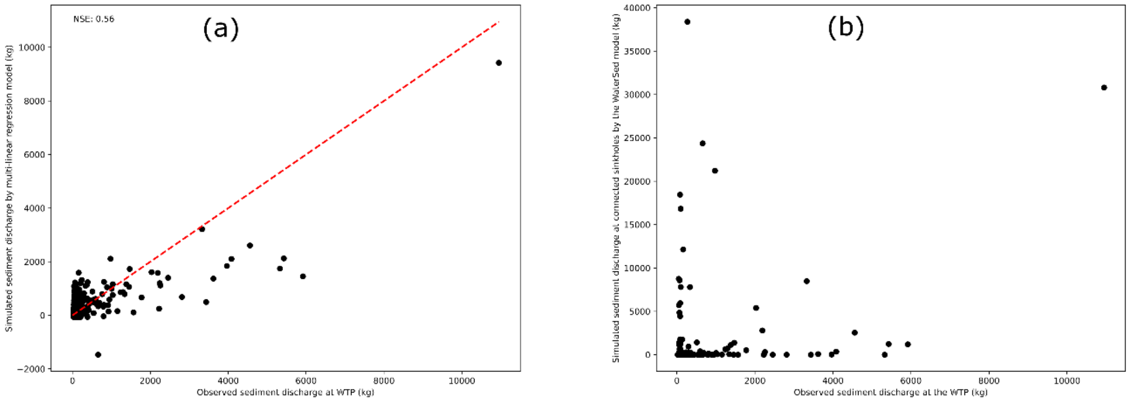

Figure S1: (a) scatter plot of sediment discharge (kg) observed at the WTP vs sediment discharge (kg) simulated by the multi-linear regression model, (b) scatter plot of sediment discharge (kg) observed at the WTP vs sediment discharge (kg) simulated at the connected sinkhole by the WaterSed model.

We also considered your remark, and did the scatter plot of the sediment discharge vs runoff simulated by the WaterSed model (Fig.S2). We can observe a high dispersion for the most extreme events. The reason is that the WaterSed model varies the erosive and runoff properties related to crops and rain characteristics.

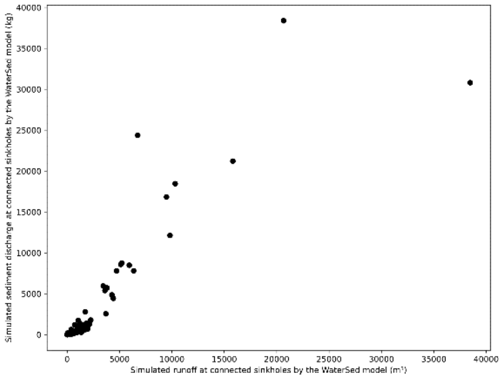

Figure S2: Scatter plot of the simulated runoff (m^3) vs simulated sediment discharge (kg) at the connected sinkholes by the WaterSed model.

I believe the part of the methodology devoted to the Deep Neural Network could be improved (see the attached file for suggestions). In particular, I would like to understand why the authors used a 3rd hidden layer with one linear node followed by an output layer which - I assume - is linear as well. I believe such a 3rd hidden layer may be redundant.

Response: Thank you for your suggestion. You were right, we did the test and the 3rd hidden layer was totally redundant. We had the same results without this hidden layer. Thus, we removed the 3rd hidden layer from the DNN. We also modified the text in corresponding section according to your suggestion and those of reviewer 2.

Line 259: "The final structure of the DNN was composed of one input layer with 4 variables, two hidden layers, and one output layer with the targeted variable. We set 40 neurons in the hidden layers as follows: 30-10. We used the rectified linear (ReLU) activation function for the two hidden layers. The optimal number of iterations was set to 15, and the batch size to 1. We used the runoff and sediment discharge simulated by the WaterSed model for the two selected hydrologic years as inputs for the DNN."

I have accidentally come across a figure very similar to figure B1 online (https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9), in an article about the training and validation of data-driven models for time series. The online article employs much of the same technical words as the authors of the paper and is not cited (or I missed it). The online article is quite good and, if it is the case that the authors used it to validate or strengthen their approach, I do not see a problem in citing it.

Response: Thank you for your suggestion. Indeed, we used the online article to strengthen our approach. Considering your suggestion, we cited the author in the methodology section and in the caption of figure B1.

Line 141: "With time series data, care must be taken when splitting the data in order to prevent data leakage (Cochrane, 2018)."

Caption Fig B1: "Month-backward chaining nested cross-validation procedure developed to test the generalization capacity of the model (modified after Cochrane (2018))".

We also added the following reference in the reference section:

Cochrane, C. (2018). Time Series Nested Cross-Validation. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

Two analyses were introduced only late in the paper and I believe the overall result could benefit from them being treated on par with the others. These are 1) the combined scenario (better farming practices and eco-engineering) and 2) the validation of the approach for extreme values resorting to the GEV distribution.

Response: Noted with thanks. Considering your suggestion, we treated the results of the combined scenario and GEV on par with the others. The methodology for the GEV was moved into the methodology section:

Line 156: "Finally, to ensure that the model does not suffer from a weakness when making simulations during extreme events, we performed an additional evaluation using the Generalized Extreme Value distribution (GEV) that is broadly applied to extreme events such as rainfall or river discharges (Carreau et al., 2013; De Michele and Avanzi, 2018). The GEV distribution was fitted to SD time series at the WTP of Radicatel. The maximum annual SD observed at the WTP was extracted for all complete hydrologic years (i.e. annual maximal blocks; n = 22) and the distribution was fitted using the package 'extRemes' of the R software (R Development Core Team, 2008; Gilleland and Katz, 2016; see Fig. C1). We compared the SD simulated at the WTP by the DNN to the values calculated by the GEV."

The results for the GEV distribution were moved to:

Line 295: "To validate the approach for extreme values, we simulated overall SD for the baseline scenario at the connected sinkholes and for the five DS using the WaterSed model. Then, we used these modelling results as inputs for the DNN model and applied it to the five DS. Secondly, we compared the results with the calculated GEV distribution (Fig. 6). The SD simulated at the WTP by the DNN model can be considered to be in good agreement with the values calculated by the GEV, even if the latter evidences a high dispersion for higher return periods. In parallel, we selected the 1% highest sediment discharge on records from October 1998 to September 2000 which represented 28% of the total sediment discharge observed at the WTP, and compared the results with simulated sediment discharge in a scatter plot. We observed a good relation between those two variables ($R^2 = 0.72$; Fig. D1), which strengthens our confidence in the model for simulating extreme events. It can be noted that, while deep learning-based methods performances to model rarely occurred events are discussed (Zhang et al., 2019), the results obtained here give confidence in the model's ability to simulate them thank to a careful selection of inputs data allowing the model to learn patterns of extreme events in the historical time series.

For the combined scenario, we removed the figure 10. The results of the combined scenario are now included in Figure 9. We modified the caption. Here is the new figure 9:

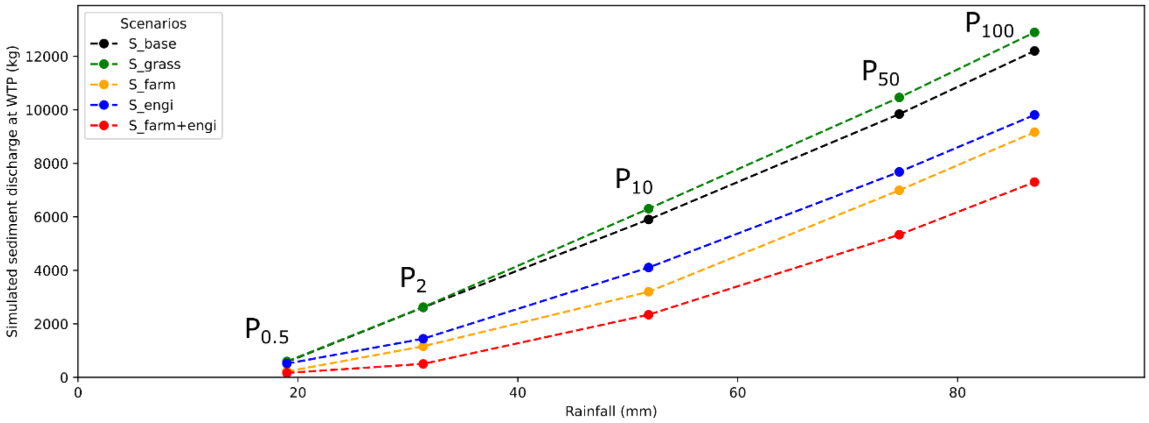

Figure 9: Simulated sediment discharge (kg) for designed storms and all scenarios at the Radicatel water treatment plant (WTP) using DNN model (S_base = baseline scenario in 2018; S_grass = 33 % of grasslands ploughed up; S_farm = +15 % infiltration capacity on 50 % of the plots; S_engi = implementation of 181 fascines and 13.1 ha of grass strips, S_farm+engi = combination of S_farm and S_engi).

We also modified the actual figure 7 and his caption to include the results of the combined scenario.

Here is the new figure 7:

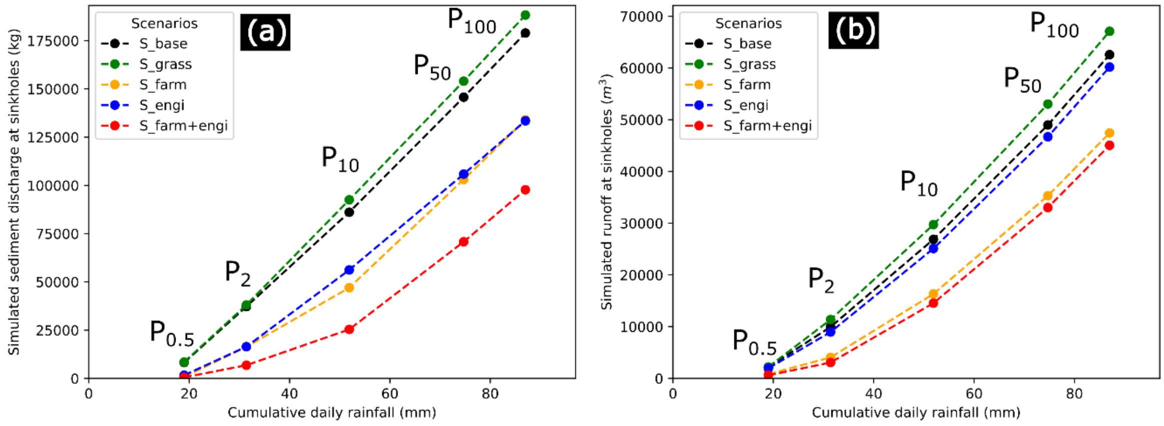

Figure 7: WaterSed modelling results according to the five designed storms for the four scenarios: Simulated (a) sediment discharge (kg) and (b) runoff (m³), extracted and summed over the 7 connected sinkholes on the Radicatel catchment. (S_base = baseline scenario in 2018; S_grass = 33 % of grasslands ploughed up; S_farm = +15 % infiltration capacity on 50 % of the plots; S_engi = implementation of 181 fascines and 13.1 ha of grass strips, S_farm+engi = combination of S_farm and S_engi).

We also made the following changes in the text according to your suggestion.

Line 19: "The sediment discharge was simulated for five designed storms under current land use and compared to four land use scenarios (baseline, ploughing up of grassland, eco-engineering, best farming practices, and coupling eco-engineering/best farming practices)."

Line 26: "Simulations made for the four land use scenarios suggested that ploughing up 33 % of grasslands would not significantly increase sediment discharge at the water treatment plant (5 % on average)."

Line 215: "Four land use change scenarios by 2050 were investigated and compared to a baseline land use scenario (2018). The scenarios were incorporated in the model in order to simulate SD variability at the WTP and evaluate their impacts. All scenarios are described below:"

Line 239: "Coupling eco-engineering and best farming practices (S_farm+engi): Both scenarios, S_farm and S_engi were combined."

Line 325: "The third scenario (S_farm) represented the adoption of good farming practices (i.e. +15 % infiltration capacity) on 50 % of the plots"

Line 330: "The fourth scenario (S_farm+engi) combined both effects of the erosion control measures and good farming practices. We observed the highest decrease in sediment discharge and runoff at connected sinkholes. The simulated values on SD ranged from 547 to 97739 kg and from 598 to 45051 m3 for runoff."

Line 352: "The third scenario (S_farm) was more efficient with simulated SD ranging from 223 to 9165 kg and a high average decrease of 43 %."

Line 355: "For the last scenario (S_farm+engi), we observed a high average decrease of 59.6%. The simulated SD values at the WTP ranged from 167 to 7298 kg. This scenario was more efficient for the 0.5, 2, and 10 years return periods (70% in average)."

Finally, to validate the model for extreme values, it would be good to see a plot similar to the ones presented, but with predictions vs observations at the water treatment plant for the highest sediment discharge events on record. This way the natural variability of the phenomenon could be accounted for.

Response: Thank you for your suggestion. We took the 1% highest sediment discharge on record which represented 28% of the total sediment discharge observed at the WTP during the two years (Fig.S3). We observed a good relation between observed and simulated extreme events (R^2 = 0.72) which strengthens our confidence to the model for simulating extreme events.

We thought that these results were important for the reader, so we added the following sentence in the manuscript, and the following figure in the Appendix (Fig.D1 in the manuscript).

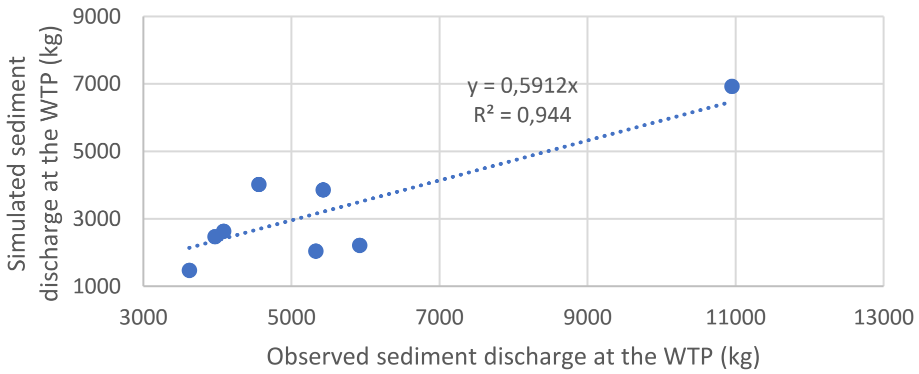

Figure S3: Scatter plot of the observed vs simulated sediment discharge by the DNN at the WTP for the 1% extreme events on records.

Line 300: "In parallel, we selected the 1% highest sediment discharge on records from October 1998 to September 2000 which represented 28% of the total sediment discharge observed at the WTP, and compared the results with simulated sediment discharge in a scatter plot. We observed a good relation between those two variables (R^2 = 0.72; Fig. D1), which strengthens our confidence in the model for simulating extreme events."

I would like to thank the authors for their efforts. Overall, notwithstanding this rather long text, I found the paper good and worthy of publication after some improvement.

Response: We thank you for your positive feedbacks.

3. Specific comments

Line 29: How is 40% achieved, being that it is more than the sum of the lower boundaries of the two individual measures (10 and 24%). I understand this is possible, but would like to know the physical explanation.

Response: Our scenario which considered best farming practices (i.e. +15% infiltration capacity) reduces sediment production on plots and peak runoff rates. In the WaterSed model, fascines are governed by a deposit function based on in-situ experimentations (Ouvry et al., 2012b). This deposition function expresses the percentage of the incoming sedimentary flow that is deposited at the fascine. It is a function of the specific incoming flow (Figure S4 below). The lower the flow rate, the greater the rate of sediment deposition. Here, we observed a cumulated effect.

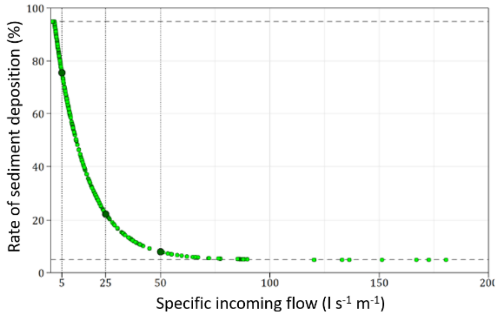

Figure S4 : Fascines rate of sediment deposition function in WaterSed model.

Line 45: Why their separate treatment is a problem should be explained

Response: Thank you for your question. We explained this point adding the following precision:

Line 46: “Several approaches have been proposed to model erosion/runoff and the karstic response induced by rainfall, but these approaches are often treated separately which does not make it possible to evaluate the impact of a change in land use on the sediment load delivered to a WTP.”

Line 48: Should be supported by references

Response: Noted with thanks. We added the following reference Verstraeten et al. (2007) in the text and in the reference section.

Line 49: “Empirical models, such as RUSLE (Renard, 1994), are frequently used because of limited data requirements but are not able to fully represent spatial and temporal dynamic of erosion processes at the catchment scale (Verstraeten et al., 2007).”

Verstraeten, G., Prosser, I.P., Fogarty, P. (2007). Predicting the spatial patterns of hillslope sediment delivery to river channels in the Murrumbidgee catchment, Australia. Journal of Hydrology, 334 (3-4), 440-454 pp. https://doi.org/10.1016/j.hydrol.2006.10.025

Line 59: I find this sentence somewhat opaque. Could the authors be more specific?

Response: Thank you for your comment. The sentence was incomplete. We rephrased the sentence as follow:

Line 62: “Based on systemic approaches, as initiated by Mangin (1984), the karst can be considered as a system able to both transform an input (rainfall) into an output (discharge) and the input-output relation can be evaluated using mathematical functions.”

Line 62: Can the "recent research" be cited?

Response: Thank you for your question. We added the following references in the text that emphasized the advantages of using data-driven techniques.

Line 64: “This approach can be considered a ‘black-box’ model to some extent, and recent research emphasized the advantages of using data-driven techniques, such as Deep Neural Networks (DNN) in similar situations (Yaseen et al., 2015; Kratzert et al., 2018; Kratzert et al., 2019b).”

2 references were added to the “references” section:

Yaseen, Z.M., El-shafie, A., Jaafar, O., Afan, H.A., Sayl, N.K. (2015). Artificial intelligence based models for stream-flow forecasting: 2000-2015. Journal of Hydrology, 530, 829-844 pp. http://dx.doi.org/10.1016/j.jhydrol.2015.10.038

Kratzert, F., Klotz, D., Herrneger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S. (2019b). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research. 10.1029/2019WR0260065

Line 67: Bluntly speaking, this is a claim that scares me and should be either removed from the paper or properly contextualized. Quite evidently, DNNs will surpass traditional statistical methods for specific problems and under specific conditions (most notably having enough data to train them with). Just stating that DNN have surpassed statistical methods (as if statistical methods were not a vast field that is also undergoing continuous evolution) is not appropriate.

Response: Noted with thanks. We considered your suggestion and simply removed the sentence.

Line 69: What do the authors mean by this?

Response: Thank you for the question. It appeared that the sentence was incomplete, so we completed it. We also moved the sentence two lines up for a better integration in the paragraph. The authors (Shen et al., 2018) meant that despite promising results with Deep Learning methods, hydrologists have not widely adopted these tools, and that DL can support a research avenue that is complementary to traditional hypotheses-driven research.

Line 69: “DNN can help providing both stronger predictive capabilities and a complementary avenue toward knowledge discovery in hydrologic sciences (Shen et al., 2018).”

Line 77: consider changing the acronym according to the previous proposal

Response: Noted with thanks. We changed the acronym DSP to DS (designed storms) according to the proposal. All changes have been made throughout the entire manuscript.

Line 88: Perhaps a paragraph break could be introduced here.

Response: Noted with thanks. A paragraph break was introduced.

Line 93: I wonder whether it is this PDF, but in the final version of the paper the quality of the figure should be improved.

Response: Noted with thanks. We thought that was a PDF problem, because all our figures were in scalable vector graphic (.svg) format exported in .png with a minimum of 300 dpi as recommended.

Line 103: Please clarify what the SAFRAN database is and where to access it.

Response: Noted with thanks. We rephrased the following sentence.

Line 188: "Cumulative daily rainfall (mm) from 1987 to 2017 was extracted from the SAFRAN database over the hydrogeological catchment of Radicatel. SAFRAN data are hydroclimatic data covering France at a resolution of 8 km on an extended Lambert-II projection and produced by MétéoFrance (Quintana-Segui et al., 2008; Vidal et al., 2010)."

Data are available on demand via MétéoFrance. We also add one reference if a reader wants more information:

Quintana-Segui, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchistéguy, L., Morel, S. (2008). Analysis of near surface atmospheric variables: validation of the SAFRAN analysis over France. Journal of Applied Meteorology and Climatology, 47, 92-107 pp. https://doi.org/10.1175/2007JAMC1636.1

Line 116: Why wasn't a year with particularly low SD "peaks" chosen as well?

Response: Thank you for the question. We conducted a statistical analysis on turbidity time series (Appendix A), to maximize the predicting capacities of the model for extreme events. It appeared that those two specific years had a very interesting distributions of values from extreme to low SD peaks. Ideally, we would have done the training on the entire time series to assess how the model reacts, but our approach required the WaterSed model's outputs as inputs in the DNN. At this time, the WaterSed model need a lot of computational resources (2 computing days for 269 events (without results verification procedure), which resulted in 60 computing days for the entire time series). We are aware that this may be an interesting perspective to evaluate the generalization capacity of the model.

Line 135: Could benefit from better rephrasing, but I find it hard to improve without risking an unwanted change to what the authors actually mean.

Response: We added some information in the sentence for the readers and a reference.

Line 120: "decision tables, adapted for the local conditions, to associate each soil surface characteristic (soil surface crusting, surface roughness and crop cover) observed in the land cover or field with a steady-state infiltration rate, a Manning's roughness coefficient, a single potential sediment concentration and a soil erodibility value (Cerdan et al., 2002a)"

Cerdan, O., Le Bissonnais, Y., Souchère, V., Martin, P., Lecomte, V., 2002. Sediment concentration in interrill flow: Interactions between soil surface conditions, vegetation and rainfall. Earth Surf. Process. Landforms 27, 193–205. https://doi.org/10.1002/esp.314

Line 142: A fragment that needs being connected to other phase or "finalized".

Response: We connected this sentence by adding a new at the beginning of the paragraph.

Line 170: "To define the soil surface characteristics needed by the WaterSed model, we computed a land cover map and a soil texture map. The land cover map was developed for 2016 by combining two national databases: the French Land Parcel Identification System (RPG) and the Soil Observatory of Upper Normandy (OSCOM). The soil texture map (three classes: clay, silt, and sand) was derived from the Regional Soil Referential (RRP) with a precision of 1:250000."

Line 158: How? Randomly?

Response: Thank you for the question. No, we did not split the data randomly to prevent data leakage. The WaterSed model considers the 48h-antecedent rainfall as input who is therefore linked to the daily cumulative rainfall. We had to keep the rain sequence. That is why we used a month-backward chaining nested cross-validation to assess the generalization capacity of the model. It allows us to maintain the rain sequence, whereas a classical cross-validation randomly split the data. We considered that a reader might ask himself the same question as you did, so we had the following precision in the text.

Line 197: "Data were split as a training set (70 %) and a test set (30 %), while respecting the chronology of daily rainfall amounts."

Line 163: Is this not part of Adam, mentioned below? Why not just to state that Adam was used?

Response: Noted with thanks. You were right, it is part of Adaptive Moment Estimation (Adam). The two sentences were redundant. We removed the following sentence: "Stochastic gradient descent was used to update the model" and just kept the sentence below:

Line 133: "The mean squared error (MSE) was chosen as loss function and the adaptive moment estimation (Adam) was adopted as the model optimization algorithm"

Line 169: Assuming that the output layer is linear too, I do not understand why there was a need to add another linear layer with just one node just before it (the last hidden layer). A ReLU function could be useful, but I fail to see what the linear layer adds to the model. The series of two linear layers with one node will probably not hurt the network's performance but, unless I am wrong, will not add anything to it neither.

Response: Noted with thanks. You were right, it did not hut the network's performance, and added anything to it neither. We did a reply to this major concern above.

Line 171: I do not think the R2, the RMSE or even the NSE need that their equations are written on the paper for the target readership of HESS.

Response: Noted with thanks. We considered the suggestion and removed the three equations on the paper. As the coefficient of determination and the Nash-Sutcliffe Efficiency were redundant as suggested by reviewer 2, we remove this part from the manuscript.

Line 150: "The performance of the model was evaluated through the Nash-Sutcliffe Efficiency (NSE) and the root mean square error (RMSE). The root mean square error (RMSE) corresponds to the standard deviation of the residuals (prediction errors). The residuals are a measure of the distance from the data points of the regression line. It evaluates how the predictions match to the observations, and values may range from no fit (+∞) to perfect fit (0) based on the relative change of the data. The Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970), indicates the model's ability to predict variables different from the mean, and gives the proportion of the initial variance accounted by the model. NSE values vary between -∞ (poor model) and 1, indicating a perfect fit between observed and predicted values."

Line 176: A very similar figure can be found in https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9 - a page that employs much of the same language as the authors did. It would be perhaps good if the authors could clarify if the page was consulted and, if so, cite it.

Response: Noted with thanks. Effectively, we used the online article to strengthen our approach. Considering your suggestion, we cited the author in the methodology section and in the caption of figure B1. More details in the comment above in the "major concerns" section.

Line 229: How do the authors define a rainfall event?

Response: Thank you for your question. This point was already defined in the methodology section, line 183.

Quote: *"For the training phase of the DNN, erosion and runoff were calculated with WaterSed model for 269 events to reduce computational efforts, considering that erosion and runoff occurs only for significant rainfall events (P > 2.5 mm/day; threshold below which no effective rainfall is generated; and runoff and sediment discharge entering connected sinkholes was set to 0)".*

Line 233: I am not familiar with WaterSed or jargon for Karstic regions. Could the authors please clarify what are positively connected sinkholes?

Response: Noted with thanks. Positively connected sinkholes are sinkholes for which dye tracing has been performed, validating the connection between surface (sinkhole) and spring. This information has been clarified in the study site section as follow:

Line 95: "According to the information system for groundwater management in Seine-Normandy (http://sigessn.brgm.fr/), hydrogeological investigations reported seven sinkholes positively connected to the springs (in-situ dye tracing already performed and confirming the connection)".

Line 250: I do not understand how this related to the "month-backward chaining nested cross-validation procedure" that was mentioned earlier.

Response: Thank you for the question. This phase corresponds to the inner loop for tuning hyperparameters, as described in the method and the appendix B.

Line 253: I believe this requires context and should be removed. 0.5 may be considered good or bad depending on the time scale of the data (e.g. monthly or daily) and the specific catchment - and that is considering only streamflow. If that is mixed with sediment loads, the suggestion that a single criterion can be used is even more risky. Personally, I would avoid comments on the quality of the results and let the reader decide whether the NSE and RMSE are good or not.

Response: Noted with thanks. We considered your suggestion and removed the sentence and the associated references (i.e. Moriasi et al. (2007) and Ritter et al. (2013)).

Line 261: Why should the predicted SD be used to quantify direct transfers? Why not only the observed values at the WTP?

Response: You are right, there was no reasons to use the simulated SD to quantify direct transfers. The observed values, and the associated recovery rates are enough to validate our results. The sentence was thus simply removed.

Line 286: Why does it not scale with rainfall? What is the physical reason that explains that the 10-year event produces more SD than the 100-year event? Is this a typo?

Response: Thank you for your question. We thought there was a misunderstanding here. The 100-year event produced more SD than the 10-year event. But in percentage, the increase was higher for the 10-year event, as you can see in the following table:

	SD_{WS} for P10	SD_{WS} for P100
Base scenario	86 102 kg	178 960
Grass scenario	92 518 (+7.4%)	188 264 (+5.2%)

To avoid any misunderstanding for the readers, we simply removed the sentence as it did not alter the content of the results.

Line 296: I find the figure ambitious but hard to read. Also, it took me some time to understand how it related to Fig. 1 (I am convinced that the sinkhole at the center of Fig. 7 is the bottom-left one of Fig. 1). I believe the readers would benefit from a less "busy" figure, perhaps with more sub-figures, each one containing less information.

Response: Thank you for your suggestion. We worked on a new figure less busy with some sub-figures as suggested.

Here is the new figure of the manuscript.

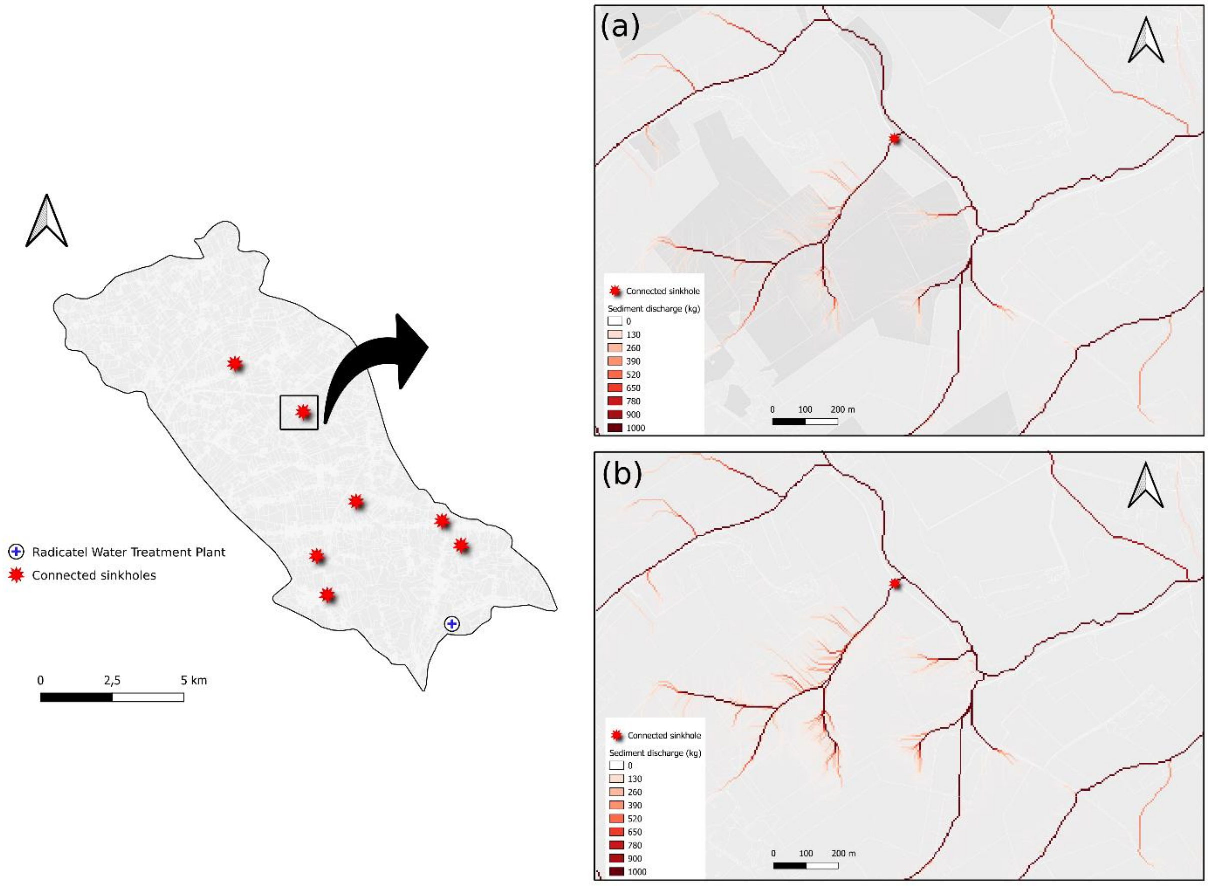

Figure 8: Mapping of flow path and sediment discharge for (a) the scenario including an increase of 15% of the infiltration capacity on selected plots (stronger grey colour) and (b) the baseline scenario, for the 10-year return period designed storm on the Radicatel catchment.

Line 348: I believe this sentence could be improved.

Response: Noted with thanks. We rephrased and improved the paragraph as follow:

Line 372: “One specific issue for DL applications in hydrological sciences, as mentioned by Sit et al. (2020) is that data provided by the authorities are dispersed and they occasionally have mismatches in temporal or spatial coverage. We thus encourage the drinking water supplier to keep turbidity records to allow the application of these data-driven models and to ensure that the time series do not have interruptions.”

We also added the following reference to the reference section:

Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. Water Science and Technology. Preprint. Doi: 10.2166/wst.2020.369

Line 305: Consider adding the return period to a second horizontal axis on top of the plot.

Response: Noted with thanks. We considered your suggestion, and we added the return period directly on the graphic to keep consistency with the other figures.

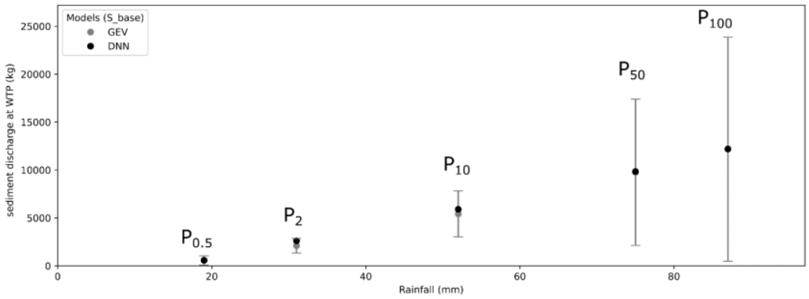

Figure 6: Simulated sediment discharge (kg) at the water treatment plant of Radicatel for the five designed storms and the land use baseline scenario by (1) Generalized Extreme Values distribution, and (2) DNN modelling. The grey bars represent the 95% confidence intervals.

Line 394: I am not sure I understand what the authors mean.

Response: Noted. Some words were missing. We meant that modelling of karstic transfers can be improved with a correct knowledge of erosion and runoff processes on the surface of the catchment. We reworded the sentence as follow:

Line 409: “This new approach does not require a knowledge of the geometry of the karstic system studied and demonstrated the value of understanding hillslope erosion and runoff processes to model underground hydro-sedimentary transfers in karstic systems.”

Line 404: Again, I am not sure about what the authors mean exactly.

Response: Noted. We wanted to discuss about connectivity and the role of grasslands spatial positions on hydro-sedimentary transfers. We reworded the sentence as follow:

Line 418: “Finally, ploughing up 33% of actual grasslands on the catchment will not significantly increase sediment discharge at the water treatment plant. However, the results might be influenced by the spatial organization of grasslands on the catchment that is a key parameter for hydro-sedimentary connectivity and hydro-sedimentary transfers to the sinkholes. In the framework of this study, we suggest conducting a specific study on this hypothesis.”

4. Suggestions

We would like to thank the reviewer 1 for the 228 suggestions in the .pdf file made to improve the quality of English. We have not listed them here, but we assure you that they all have been considered.

Reply letter to RC2

Authors: Edouard Patault, Valentin Landemaine, Jérôme Ledun, Arnaud Soulignac, Matthieu Fournier, Jean-François Ouvry, Olivier Cerdan, Benoit Laignel

Manuscript n°: HESS-2020-363

Please find in "black" the reviewer's comments, in 'blue' our replies, in 'green' the modification in the revised manuscript.

1. General comments and suggestions

The proposed article presents two topics: the first one is the realization of a deep model to estimate the Radicatel solid transport at the spring (or in boreholes?). The second is to generate different shallow solid transport scenarios from agricultural hypotheses. I personally have no knowledge of agriculture to really appreciate the interest of this approach. However, I have some doubts about the originality of the results. It seems obvious that if we want to avoid turbidity in the springs, it is better to avoid generating it on the shallow water. Concerning the quantification brought by the cascade of models carried out (it is not a coupling), I have serious doubts about the robustness of the numerical values, because no estimation of uncertainty is carried out throughout the article, and because the description of the deep model design is not sufficiently accurate to assess its quality. Overall, the article is confused, the description of the data and of the processing carried out is scattered everywhere and it is difficult to find relevant information.

Response: We thank the reviewer 2 for sharing his concerns and the constructive review which clearly improve the quality of the manuscript. In this reply, we addressed a response to each major concerns and specific comments. We have no doubt about the originality of the results. As we said to reviewer 1, we agree that the approach might not be a large breakthrough, consisting in the cascade modelling of two existing modelling tools. However, to the best of our knowledge, this approach has never been proposed in the literature. It is interesting because physically based models are generally used to study sub-surface hydro-sedimentary transfers in karstic regions. But these models require 3-D information on the physical characteristics of the studied karst system which are often not available and limit the model's applicability. This approach allows to assess the impact of multiple land use scenarios on the sediment discharge observed at a water treatment plant. We are confident that our article will have a positive impact for all researchers, drinking water suppliers, and land use planners over the entire North-European loess belt which is impacted by the same erosion processes in a similar environment. One other main contribution of this paper is that the proposed tool is integrated into a decision support system for land use planning. As mentioned by Sit et al. (2020) in their review, is that few DL applications in the water literature included decision-making components. This lack of attention to DL in service of decision-making in water academia presents an opportunity.

The robustness of the numerical values are supported by field validations, statistical analysis, or extension of the analysis like we did for the extreme events in order to verify the applicability of the model over a wide range of conditions.

The description of the deep model has been upgraded up to international standards as you recommended. And the structure of the article has been modified to follow your recommendations.

Please, find more details in the responses to each specific comment.

My recommendation is that the paper cannot be published in the current state because too many improvements must be done. I regret that, because the subject is interesting and challenging.

Suggested improvements are: (i) adopt a simpler and more efficient outline; (ii) given the small amount of data, use a shallow model, benchmark with a linear model; (iii) better present the model design and the model actually used in a very precise way (list of inputs, very precise architecture); (iv) do not shift the input series in relation to the output series: neural networks know how to manage this; (v) make a statistical description of the data in order to understand why the results on the test set are better than the training one. This, moreover, allows us to prejudge bad results on another test set. Other general remarks: Words have a signification: it is not really a coupling of models: one feed the other. Goal is focused on operational needs, but operational needs are not accurately described. Very few details on how the solid flow is obtained. What are the assumptions and the model considered? It is also written that "This new approach can be easily implemented". Applying a deep model is anything but easy. This may explain inconsistencies and poor explanations found in the paper.

Response: Thank you for your suggestions. We considered the suggested improvements. The article now adopts a more efficient outline (i.e. "methodolody" and "data" sections). We did a benchmark with a multi-linear model in the reply to show that is outperformed by the DNN. We now present the model design in a more precise way as suggested. We explained in the reply why it was more interesting to shift input series in relation to the output series. We now used the term "cascade modelling" instead of "coupled modelling", and the term "simulation" instead of "prediction" which are more appropriate. Operational needs are described in a more precise way in the introduction. And we added some details on the solid flow. Please find below more details with the responses to each specific comment.

2. Specific comments

Abstract

*P*1 L15 "and they can be seen as a black-box due to the non-linearity of the processes generating sediment discharge. Âz there is no straightforward link between the black ˙ box property and the non-linearity of the relation. Black-box means "unknown" when nonlinear means frequently "difficult".
Response: Noted with thanks. As also suggested by reviewer 1, the sentence was modified as follow:

Line 14: "Karstic environments are particularly challenging as they face a lack of accurate physical description for the modelling process, and they can be particularly complex to predict due to the non-linearity of the processes generating sediment discharge."

P1L23; is water extracted at sinkhole or spring? It could be interesting improving the consistency to enhance the impact of the abstract.
Response: The raw water is extracted at the spring, which is fed in part by fast transfers via the sinkholes.

Introduction

P1LL 39-43. Could you please be more accurate? It is not said if the cost is linked to sediments or to other dissolved pollutants.

Response: Noted with thanks. The cost is linked to sediments in raw water. The sentence was modified as follow:

Line 40: "Upper-Normandy (France) is particularly affected, and the economic cost linked to the restrictions on the use of drinking water due to excessive sediment inputs in raw water was estimated at €5 million between 1992 and 2018 (Patault et al., 2019)"

P2L66 “As stated by Shen (2018), DNN have now surpassed traditional statistical methods”. This sentence is too simple. It has no meaning by itself without the definition of “traditional” statistical methods, and the definition of the limitations. For example it is false for linear problems or for nonlinear relations in general. Only specific nonlinear relations need deep models. The drawback of deep models is that they need extensive database. Are they really relevant with highly noised data? This question could be discussed in the light of a comparison with the results of a linear model.

Response: Noted with thanks. The reviewer 1 suggested to remove the sentence. The sentence had been removed from the manuscript.

The reviewer 1 also suggested to benchmark with a multi-linear model, please find below our reply to reviewer 1.

Response: Noted with thanks. Your concerns are legitimate but before using a DNN in this study, we tried a simpler approach with a multi-linear regression model. The results of the simulation with the regression model are presented (Fig.S1a). We observed that the model was unable to correctly simulate sediment discharge, we observed a high dispersion of the values specifically for the most important events. The results were not surprising considering that the karstic system is complex, and the link between the sediment input simulated at the connected sinkholes by the WaterSed model cannot be linked to the sediment discharge at the spring with a simple linear response (Fig.S2b). The response is non-linear considering multiple processes like fast infiltration, matrix infiltration, underground storage, purge, etc. So, we benchmarked a multi-linear regression model vs the DL model with the same inputs. It appears that the DNN outperformed the multi-linear regression model because the DNN was able to consider the non-linearity of the system.

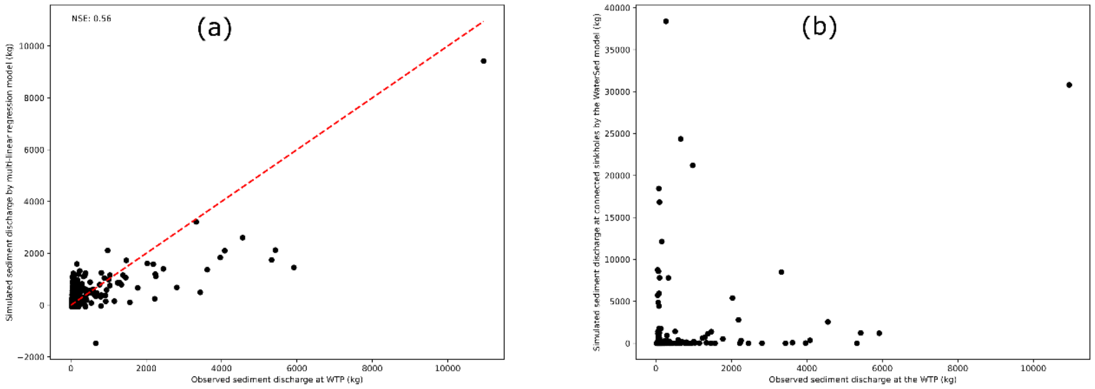

Figure S1: (a) scatter plot of sediment discharge (kg) observed at the WTP vs sediment discharge (kg) simulated by the multi-linear regression model, (b) scatter plot of sediment discharge (kg) observed at the WTP vs sediment discharge (kg) simulated at the connected sinkhole by the WaterSed model.

P2-L72-78. The presentation of the goals and methods is chaotic, please could you present accurately what is the goal: prediction of which variable, where and at what time-step, with or without rains?, . .

Response: Noted with thanks. We considered your suggestions and reworded the goal of the study in the introduction. We considered your other concerns and divided the manuscript with a specific section for “methodology” and one for “data”.

Line 73: “The main objectives of this study were: (i) to develop a cascade modelling approach able to simulate hydro-sedimentary transfer at a WTP for specific daily rainfall events, and (ii) to evaluate the impact of different land use scenarios on the SD variability. This study was conducted in the Radicatel hydrogeological catchment (Normandy, France) where spring water is extracted to a WTP. We benefited from the use of an existing expert-based GIS model (WaterSed), developed and successfully applied on the studied area, to simulate the impact of land use management on hydro-sedimentary transfers to connected sinkholes. Rainfall events characteristics and WaterSed outputs were then used as input for a data-driven model (i.e. DNN) to simulate SD at the WTP. The cascade modelling approach was applied to multiple design storms (DS) under different scenarios in order to simulate hydro-sedimentary transfer in the hydrogeological catchment and evaluate the efficiency of different land use management strategies."

One remark about vocabulary. Usually the words prediction or forecast are used with actual data. When data correspond to scenario, it is not a prediction, maybe a prospective?

Response: Noted with thanks. We changed the title of the section "Scenario predictions" to "Prospective Analysis" and we considered that the hydro-sedimentary transfers are "simulated" instead of "predicted". The modifications have been applied in the entire manuscript.

2. Study site

It is not clear if the cited catchment (106 km2) is the underground basin, or the shallow basin. Both basins can be different in karst context. Could you also clarify the notion of "positively connected"? By tracing tests, by signal analysis?

Response: The cited catchment is the underground basin from which drinking water is abstracted. To avoid any misunderstanding, we added the term "hydrogeological" to catchment in the abstract, introduction, and method sections. As also suggested by reviewer 1, the notion of positively connected had been clarified as follow:

Line 95: "According to the information system for groundwater management in Seine-Normandy (http://sigessn.brgm.fr/), hydrogeological investigations reported seven sinkholes positively connected to the springs (in-situ dye tracing already performed and confirming the connection)".

L71, one cited goal is to study "hillslopes erosion processes into karstic transfer", but elevation is not provided and the plateau seems very flat. Is there an explanation?

Response: The elevation has been calculated from the 5m resolution DEM available and is now provided in the study site section of the manuscript. Even if the plateau seems flat, it's not, and in the Normandy region we observed erosion processes on low slopes on the loamy soils due to soil crusting on agricultural plots.

Line 88: "The elevation ranges from 138 to 0 m and the median slope is 5.9%."

Finally the presentation of the study site could be improved. The same apply to the hydroclimatic data. Not all types of data are presented. This section is poor and must include information provided in section 3.1 "Data handling". A comprehensive description of data is important for data-based models.
Response: We considered your suggestion and added some information to the section "2.Study site". We also divided the old section 3 into two new sections: "3.Methodology" and "4.Data". We also added some information on the hydroclimatic data as suggested by reviewer 1. Please see changes in the entire manuscript.

3 Methodology

L113 what did you mean by "training limits"; is it training set?

Response: Yes, it is.

I have a concern with the word "coupling", or coupled models. Coupling don't means preprocessing or cascading processes; coupling means that processes and their attached data depend each one from the other, as for example in coupled differential equations. If my comprehension is good it is not the case in the paper. Please clarify this point.

Response: We considered your suggestion and deleted the word “coupling”. We now used the term “cascade modelling” which is more appropriate. Please see changes in the entire manuscript and in the title.

New title: “Simulating Sediment Discharge at Water Treatment Plants Under Different Land Use Scenarios Using Cascade Modelling with an Expert-Based GIS Model and a Deep Neural Network”

L120, I do not understand this sentence: “In accordance with previous studies by Masséi et al. (2006) and Hanin (2011) in karstic environment in Normandy, a lag of 1 day was applied to the SD time series to properly match the rainfall input”. Have you shown before that the ANN was unable to calculate the good lag? How is managed this artificial delay in the following results?

Response: We performed a classical trial and error procedure to find an optimal model design. During this phase, we tested different combination of inputs data (ex: which rainfall parameter to include, etc.) and model structures (number of neurons and hidden layers, hyperparameters, etc.). We knew that the response of sediment transfers from sinkholes to the spring had a delay (measured by dye-tracing, two PhD thesis were performed on that study site between 2008-2014 (Hanin, 2011; Chédeville, 2015)). So, we hypothesized that the consideration of that lag in the inputs should improve the performance of our model, and it was the case. This delay is managed by simply matching the input of the day ‘t’ to the output of day ‘t+1’.

L 117, please could you explain what is the difference between calibration and training, and what is a reference period? In my opinion it is only training.

Response: Thank you for the suggestion. There was no difference between calibration and training, but there was a misunderstanding in the sentence. The two hydrologic years were selected as the complete set (training + test) for the model. And data were split as a training set (70 %) and a test set (30 %), while respecting the chronology of daily rainfall amounts. The sentence has been corrected in the manuscript.

Line 197: “Data were split as a training set (70 %) and a test set (30 %), while respecting the chronology of daily rainfall amounts.”

L132 “burn the stream network”?

Response: It corresponds to the modification of the DEM to ensures a steady downstream gradient according to the observed hydrographic network. This procedure is part of a SAGA-GIS module (http://www.saga-gis.org/saga_tool_doc/2.2.3/ta_preprocessor_6.html). The sentence has been reworded in the manuscript.

Line 118: “stream network (used to modify the DEM to ensure a steady downstream gradient according to the observed hydrographic network) and river width”

The section “Erosion and runoff modelling” is badly organised: the description of the data should be put in a "data" chapter, the method in a "method" section, and its limitations should be given. Here everything is mixed up. You get lost in the article.

Response: Thank you for the suggestion. We put the methodology in the new section “3.Methodology” and the description of the data in the new section “4. Data”. The main limitation is that the model is non-dynamic, but I am not sure this respond to your concern. Can your concern be more specific?

3.3 DNN Configuration

First of all, as I understand it, the amount of data used for training is only 751 examples. This seems very little for a deep model, which necessarily includes a lot of parameters to adjust. Shen et al 2018 that is cited in the paper wrote also: “Despite the comparisons between DL and nondeep machine learning, this author would strongly advise against applying DL nondiscriminatively. The earlier generation methods can be highly valuable for their respective problems and situations, especially when there are limited data or relatively homogeneous data. As shown in studies, for small data sets, DL could be at a disadvantage compared to models with stronger structural assumptions.”

Response: Thank you for sharing your concern. Please see above our reply to reviewer 1 concerning the benchmark with a multi-linear model.

I am sorry, but the presentation of the model is not up to international standards. It is necessary to indicate precisely how the data and related time windows are selected, how overfitting is avoided (or underfitting, given the small number of examples), and accurately what structure is obtained after these steps. Are there any regularization methods used?

Response: Noted with thanks. The selection of the data was already indicated in the manuscript (even with a statistical description of the time series in the Appendix). We also mentioned how we split the data. We now added a sentence to show that we respected the chronology of the daily rainfall amounts to prevent data leakage as suggested by reviewer 1. We also totally modified the Methodology section to add some information for the readers, and up to international standards. In this study we used two regularization methods: (i) cross-validation and (ii) early stopping. You were right, the latter was missing in the manuscript and we now added this information. The structure obtained after these steps was already described in the methodology section, but the location was inadequate and the paragraph was relocated to the beginning of the results section (5.2.2).

Line 197: “Data were split as a training set (70 %) and a test set (30 %), while respecting the chronology of daily rainfall amounts.”

Line 126: “In this study, a multi-layer feed forward network (DNN) was built under Python version 3.6 using the high-level API Keras (Chollet, 2015) and Tensorflow (Abadi et al., 2016) as a background engine. A multilayer feedforward neural network is an interconnection of perceptrons in which data and calculations flow in a single direction, from the input data to the outputs. The number of layers in a neural network is the number of layers of perceptrons. The number of hidden layers and perceptrons depends on the characteristics of the input data, and there is no specific rule for selecting these parameters (Le et al., 2019). Neural networks are adjusted during a training stage when the parameters are calculated iteratively a gradient descent that seeks to minimize a mean squared error. For efficient learning, input variables were rescaled between 0 and 1 using normalization and re-transformed for the simulations using the normalization parameters. The mean squared error (MSE) was chosen as loss function and the adaptive moment estimation (Adam) was adopted as the model optimization algorithm. The quality of the model is obtained after training and evaluated on a test set. The model’s ability on the test set is called generalization and this can be affected by overfitting on the training stage. To avoid overfitting and increase model robustness, we used two regularization methods in this study: (i) early stopping and (ii) cross-validation. Early stopping is an efficient regularization method that prevents the model from overfitting (Sjörberg and Ljung, 1992). We visually monitored the learning curves and stopped the training as soon as the validation error reached a minimum. Then, we rolled back the model parameters to the point where the validation error was at the minimum. The choice of the training/test set remains arbitrary and may need to be evaluated to compute a robust estimate of model error. With time series data, care must be taken when splitting the data in order to prevent data leakage (Cochrane, 2018). To address this issue, we adapted a month-backward chaining nested cross-validation procedure, which provides an almost unbiased estimate of the true error (Varma and Simon, 2006). The procedure contains an inner loop for parameter tuning, and an outer loop for error estimation (see Fig. B1). The parameters that minimize error are chosen on the inner loop and we add an outer loop, which splits the data into multiple training/test sets. Then, the error is averaged on each split to evaluate the overall performance of the model. The optimum structure and configuration of the network (model design) was found by a classical trial and error procedure (training-evaluation process through optimization of errors; Ortiz-Rodriguez et al., 2013)."

The following references were added in the manuscript

Cochrane, C. (2018). Time Series Nested Cross-Validation. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

Sjörberg, J. and Ljung, L. (1992). Overtraining, regularization, and searching for minimum in neural networks. IFAC Adaptative Systems in Control and Signal Processing, Grenoble, France, 1992.

3.4 Performance evaluation

L179 There is a confusion between the coefficient of determination and the linear correlation coefficient (eq 1). The coefficient of determination is equal to the Nash efficiency. Please correct this error.

Response: Thanks for your awareness. The error was corrected in the text. To avoid redundancy, we deleted the part on the coefficient of determination.

The use of the word "prediction" also needs to be discussed. Obviously the results represent what might happen in the future; but the models are fed with data that are also into the future. Actually the model only makes an estimate. If the model performs a prediction (effective anticipation from inputs in the past or present) then quality criteria specific to the prediction should be used, such as the persistence criterion. This is not the case

Response: Noted with thanks. We considered your suggestion and now used the word "simulation" in the entire manuscript. We also renamed the section 5.3 as "prospective analysis". We think that these words are more appropriate.

3.5 Designed storm projects and land use scenarios : to put also in a data section.

Response: We considered your suggestion and put the old section 3.5 into the new section "4.Data"

4.2. DNN: Calibration and Generalization

The monthly-backward chaining nested cross-validation procedure must be defined more clearly and more accurately as this process is critical.

Response: Noted, but we think that the description is sufficient. We added an entire paragraph and an appendix in the original draft. And as suggested by reviewer 1, we now added a new reference for the readers. Here is the response to reviewer 1 :

> *I have accidentally come across a figure very similar to figure B1 online (https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9), in an article about the training and validation of data-driven models for time series. The online article employs much of the same technical words as the authors of the paper and is not cited (or I missed it). The online article is quite good and, if it is the case that the authors used it to validate or strengthen their approach, I do not see a problem in citing it.*

Response: Thank you for your suggestion. Indeed, we used the online article to strengthen our approach. Considering your suggestion, we cited the author in the methodology section and in the caption of figure B1.

Line 141: "With time series data, care must be taken when splitting the data in order to prevent data leakage (Cochrane, 2018)."

Caption Fig B1: "Month-backward chaining nested cross-validation procedure developed to test the generalization capacity of the model (modified after Cochrane (2018))".

We also added the following reference in the reference section:

Cochrane, C. (2018). Time Series Nested Cross-Validation. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

I definitively not understand the sentence "Predicted values of runoff and sediment discharge were extracted over the connected sinkholes and summed to be used as inputs . . .", please correct it: how can future data be "extracted" from a sinkhole? An how can we sum runoff and sediment discharge?

Response: The WaterSed model is a distributed model (resolution 5m). The sentence means that the WaterSed model simulated runoff and sediment discharge for each event (from October 1998 to September 2000) and for each one of the seven sinkholes on the underground basin. The seven sinkholes are connected to the spring (positive dye tracing). So, we summed for each event the values predicted to each sinkhole to consider a unique contribution to the karstic spring. Runoff and sediment discharge were summed independently. This was not future data that were extracted, here we reconstructed past erosion and runoff from Oct-98 to Sep-2000 on the underground basin. To avoid any misunderstanding, we rephrased two sentences as follow:

Line 246: "For each event, WaterSed outputs (i.e. runoff (R_{WS}; m^3) and sediment discharge (SD_{WS}; kg d^{-1}) values) were independently extracted over the connected sinkholes and summed to consider a unique contribution from the 7 sinkholes to the spring."

Line 261: "We used the runoff and sediment discharge predicted by the WaterSed model for the two selected hydrologic years as inputs for the DNN."

What is the accuracy of the turbidity/Solid-transport relation?

Response: We now added the accuracy of the turbidity/soli-transport relation directly in the manuscript.

Line 199: "At the Radicatel WTP, the mean pumping flow rate is estimated to 19733 m3 d-1, and the volume of water pumped in 2018 was approximately 7.2 million m3. SD (kg d-1) was assessed considering turbidity time series, mean pumping flow rate at the Radicatel WTP and the relation between turbidity and sediment concentration resulting from field investigations ([mg L-1] = 0.96*[NTU]; $R^2 = 0.97$; Hanin, 2011)."

L 268. What does mean the sentence: "The modelling results were less efficient than for the complete dataset but overall satisfactory"

Response: The sentence meant that the performances of the model were better on the inner loop (i.e. complete data = 731 events) than for the 12 outer loops (Fig. 5 and Fig. B1). To enhance understanding, we reworded the paragraph as follow:

Line 282: "This procedure removed one month from the initial dataset (i.e. 731 events) and was repeated twelve times, while keeping at least one full hydrologic year as input for the modelling. The modelling results were more efficient for the inner loop (Fig. 4) than for the outer loops (Fig. 5 and Fig. B1). Overall, the median NSE value for the training and the test sets for the outer loops was above 0.5."

Results suggest that test data are better represented than training data. This suggest that the good quality obtained on test set will not be generalizable to the part of data used in training test. Results seem to be overestimated.

Response: We agree with your remark that test data were slightly better represented than training data. We were totally transparent and already mentioned this point in the manuscript. We faced a classical bias-variance dilemma. Nevertheless, the slight overestimation was controlled because during training stage we used regularization methods as mentioned above.

5 Discussion

Is it possible to describe the Generalized Extreme Value Distribution (GEV) in the material and method section?

Response: Noted with thanks. This point was also suggested by reviewer 1. The GEV description was moved to the material and method section.

L 345: "Even if it is well known that deep learning-based methods may results in weak performance for extreme events (Zhang et al., 2019)". Then I no longer understand the coherence of the article: why did you use this method on extreme events?

Response: We assure you that the article is coherent. In fact, the sentence was not properly contextualized. Zhang et al. (2019) suggested that data imbalance is an important issue for which DL can be unable to model extremes events. Same authors suggested the concept of memory network, the main idea is to memorize extreme events in historical data. To this end, that's why we have conducted the statistical analysis on the 22-years turbidity time series at the water treatment plant (Appendix, Fig.A1). In parallel, as suggested by reviewer 1, we compared observed versus simulated sediment discharge for the 1% highest sediment discharge on records to strengthen our approach. To avoid any misunderstanding, we reworded the paragraph as follow:

Line 300: "In parallel, we selected the 1% highest sediment discharge on records from October 1998 to September 2000 which represented 28% of the total sediment discharge observed at the WTP, and compared the results with simulated sediment discharge in a scatter plot. We observed a good relation between those two variables ($R^2 = 0.72$; Fig. D1), which strengthens our confidence in the model for simulating extreme events. It can be noted that, while deep learning-based methods performances to model rarely occurred events are discussed (Zhang et al., 2019), the results obtained here give confidence in the model's ability to simulate them thank to a careful selection of inputs data allowing the model to learn patterns of extreme events in the historical time series."

6 Conclusion

L393. "This new approach can be easily implemented". Applying a deep model is anything but easy
Response: You are right. We removed this part of the sentence in the manuscript.

Line 409: "This new approach does not require a knowledge of the geometry of the karstic system studied and demonstrated the value of understanding hillslope erosion and runoff processes to model hydro-sedimentary in karstic systems."