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.

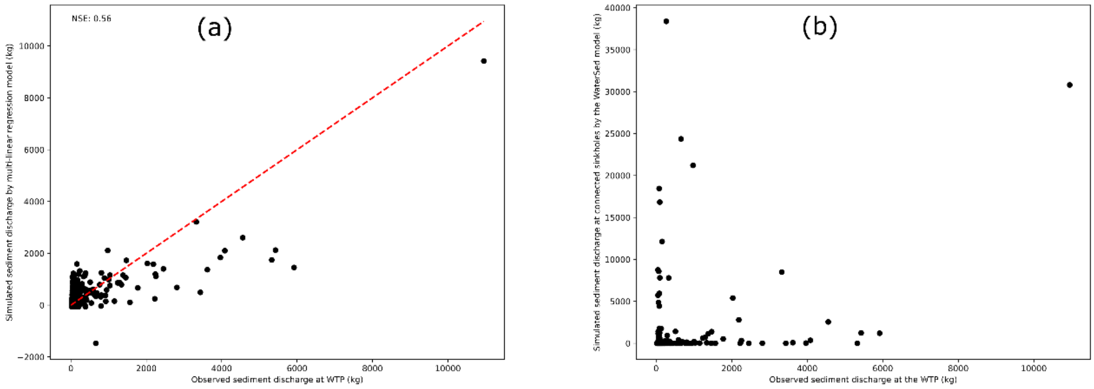

Figure S1: (a) scatter plot of sediment discharge (kg) observed at the WTP vs sediment discharge (kg) simulated by the multi-linear regression model, (b) scatter plot of sediment discharge (kg) observed at the WTP vs sediment discharge (kg) simulated at the connected sinkhole by the WaterSed model.

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.”