Answers to Reviewer#2 comments

We thank Reveiwer#2 for his/her comments. The reviewer comments appear in black and the answers appear in blue.

1/ The paper focuses on the use and evaluation of the Indicator of Intense Pluvial Runoff (IRIP) method. The method is used to provide susceptibility maps for a railway line in Northern France. The paper is well written and organized. The topic is also of high interest. However, one major issue in the methodology may undermine the results of the entire work.

We thank Reviewer#2 for his.her encouraging comments regarding the interest of our work. We also thank him.her for his.her suggestions regarding the evaluation methodology. We provide a detailed answer below and how we propose to take the suggestions into account in the revised version of the manuscript.

2/ Moreover, the authors should be very careful in not using the same figures and same wording used in their previous works.

See answer to comment 8/ below.

3/ Major comments:

This is the critical issue: In Step 4 it is written that "If an area classified at risk has a specific supervision measure or mitigation structures have been built, it is moved from 'false alarm' to 'hit' as the implementation of mitigation measures means that the area was indeed at risk, but that no impact as recorded due to the efficiency of the mitigation measure." Step 4 brings a significant a bias in the evaluation, which can be seen as a unidirectional attempt to improve model performance. If authors could like to use an approach taking into account mitigation measures, they should also do the opposite: "If an area not classified at risk has a measure or structure, and no impact was recorded, it should be moved from 'Correct negative' to 'miss'", or, at least, according to section 2.4, if the area was in the "bad weather tour".

We thank Reviewer#2 for raising this point. We acknowledge that this point is worth being considered in step 4 of the evaluation methodology, provided that the proxy data can really be related to runoff-related risks. If this is the case, we agree with Reviewer#2 that step 4 of the methodology must be modified and that "If an area not classified at risk has a measure or structure, and no impact was recorded, it should be moved from "Correct negative" to "miss'".

Concerning the mitigations measures considered in our case study, the following comments can be made: 1/ Contrarily to runoff-related impacts, hydraulic works and "bad weather tours" only indicate a potential risk, not a proven risk (i.e. a risk that already happened); 2/ The use of mitigations measures as proxy data must be done with caution. Indeed, the construction of hydraulic works or the design of tours take into account not only the hazard parameter, but also the vulnerability and the criticality of the stake. A "bad weather tour" is preferably designed on a section that is critical regarding the train traffic management, or on sections with known structural weaknesses. "Bad weather tours" are not precise, they often involve long linear of the railway, and the whole sections of the tours may not be relevant to runoff risk. This is why we only used them as explanatory factors of false alarms, and not as elements proving the existence of a risk (implying to move "correct negative" to "miss" if an area was not classified at risk but had a hydraulic work or was part of the "bad weather tour").

Nevertheless, we propose to modify the description of step 4 as follows and to add the above discussion to the discussion section.

“This fourth step is necessary to properly take into account the fact that, if an area is at risk, the stakeholder may have taken mitigation measures that may explain the absence of observed impact. Such mitigation measures are therefore likely to explain a certain amount of false alarms. For instance, mitigation structures can be protection to buildings, retention basins, hydraulic works crossing below transport infrastructures, etc… They can also be resilience actions like a reinforced supervision in case of high rainfall amount warning. If an area classified at risk has a specific supervision measure or mitigation structures have been built, it is moved from ‘false alarm’ to ‘hit’ as the implementation of mitigation measures means that the area was indeed at risk, but that no impact was recorded due to the efficiency of the mitigation measure. This step will be referred to Step 4.1 in the following. If mitigation measures can be considered as a reliable source of information regarding runoff risk, the section must also be moved from “correct negative” to “miss” if a mitigation measure is present and the area was not classified at risk (this step will be referred as Step 4.2 in the following. The performance measures are then recomputed, based on the modified contingency tables of Step 4.1 or Step 4.2 is the latter is relevant.
After these four steps, the final values of the quantitative performance measures are obtained, characterizing the performance of the IRIP mapping model. “

Furthermore, in order to investigate the impact of considering Reviewer#2’ suggestion in the evaluation, we performed additional computations of the evaluation criteria, taking into account Reviewer#2 suggestion, to see how it modifies the computed performance criteria of the IRIP model. The reference results are those of column 2 in Table 5, where only the vulnerability of the railway is taken into account in the evaluation. We computed the criteria by taking into account mitigation measures as follows: “Stet 4.1”: moving from ‘false alarm’ to ‘hit’ the sections where a mitigation measure is present but no impact was recorded (present version of the paper), and “Step 4.2”: moving from ‘false alarm’ to ‘hit’ the sections where a mitigation measure is present but no impact was recorded AND moving from “Correct negative” to “miss’” the sections where a mitigation measure is present and the section was not tagged at risk by the IRIP model (Reviewer#2 suggestion). Mitigation measures were accounted for as follows and the results are reported in Table 1 below:

- By taking into account hydraulic works only
- By taking into account “bad weather tours” only
- By taking into account both hydraulic works and “bad weather tours”.

When comparing the results obtained in Step 4.1 and Step 4.2, the results of Table 1 show that, in Step 4.2, when only hydraulic works are taken into account, only three additional sections are moved from “correct negative” to “miss”, leading to a small decrease of POD. When only “bad weather tours” are considered as mitigation measures in Step 4.2, 9 additional sections are moved from “correct negative” to “miss”, leading to a decrease of POD from 94% to 84%. Finally, when hydraulic works and “bad weather tours” are both taken into account, POD decreases from 96% to 86%. FAR remains unchanged when moving from Step 4.1 to Step 4.2.
When comparing results of Step 4.2 with the reference results (without accounting for mitigation measures), Table 1 shows that when considering hydraulic works only, POD slightly decreases whereas FAR is greatly improved. When considering “bad weather tours” only, POD decreases significantly (from 93% to 84%) as the number of “miss” increases. Nevertheless, FAR is improved from 58% to 49%. Finally, when both hydraulic works and “bad weather tours” are accounted for, POD decreases from 96 to 86% but FAR is improved from 58 to 28%.
Even when modifying the evaluation methodology according to Step 4.2, the performance of the IRIP models remains satisfactory (see also answer to comment 6/).

Table 1: Performance criteria assessing the predictive power of the IRIP model in identifying sections with a proven risk using different methods to take into account mitigation measures. In columns 3 to 5, the figures correspond to Step 4.2 (in parenthesis to Step 4.1, i.e. the first version of this paper).

	Reference IRIP when taking vulnerability into account but not mitigation measures	IRIP when taking vulnerability and hydraulic works into account	IRIP when taking vulnerability and "bad weather tours" into account	IRIP when taking vulnerability hydraulic works and "bad weather tours" into account
Number of 'Hit'	55	85 (85)	67 (67)	95 (95)
Number of 'False Alarm'	77	47 (47)	65 (65)	37 (37)
Number of 'Correct Negative'	46	43 (46)	37 (46)	35 (46)
Number of 'Miss'	4	7 (4)	13 (4)	15 (4)
Probability of Detection: POD (%)	93	92 (96)	84 (94)	86 (96)
False Alarm Ratio: FAR (%)	58	36 (36)	49 (49)	28 (28)
χ^2	19	36.8 (46.1)	9 (27.9)	26.7 (59.8)

Table 2: Performance criteria assessing the predictive power of the presence of hydraulic works or "bad weather tours" in identifying sections with a proven risk using different methods to take into account mitigation measures

	Impacts explained by hydraulic works	Impacts explained by "bad weather tours"	Impacts explained by hydraulic works or "bad weather tours"
Number of 'Hit'	22	29	40
Number of 'False Alarm'	33	21	51
Number of 'Correct Negative'	90	102	72
Number of 'Miss'	37	30	19
Probability of Detection: POD (%)	37	49	68
False Alarm Ratio: FAR (%)	60	42	56
χ^2	2 (not significant)	20.6	11.1

Furthermore, we also computed the performance criteria to assess the predictive power of, respectively, hydraulic works only (Table 2, column 2); "bad weather tours" only (Table 2, column 3); and a combination of both sources of information (hydraulic works OR "bad weather tours") (Table 2, column 4) in explaining the recorded runoff-related impacts. The contingency tables presented in Table 2 were computed assuming that a section was at risk if a hydraulic work or a "bad weather tour" was present in this section (i.e. assuming that they could be considered as proven risk). The results show that the hypothesis that hydraulic works and runoff-related risks are independent cannot be rejected (χ^2 not significant), i.e. hydraulic works have low predictive power about the occurrence of a risk. On the other hand, "bad weather tours" have a predictive power but lower values of POD (49%) than the IRIP model without mitigation measures (93%) (comparison of column 3 of Table 2 and column 2 of Table 1). The FAR value is lower than for the IRIP model without mitigation measures (42% as compared to 58%) but is not so different. Finally, when the presence of hydraulic works and "bad weather tours" are combined (Table 2, column 4), the POD increases to 68% as compared to considering hydraulic works only or "bad weather tours" only. The FAR is 56%, an intermediate value between the one of hydraulic work only (60%) and "bad weather tour" only (42%). In any case, the predictive power of the IRIP model is higher than when considering hydraulic works or "bad weather tours" as proxy for the risk of intense runoff. The results presented in Table 2 also highlight that the "bad weather tour" is a more reliable proxy data for runoff related risk than hydraulic works.
We propose to add these elements to the discussion section.

5/ Since this may worsen performance, maybe the authors should either create a different model for areas with mitigation measures, or entirely remove them from the evaluation.
In our case study, this solution would have the drawback of excluding a large part of the runoff-related impacts from the analysis (22 impacts out of 59 also have a hydraulic works; 29 impacts out of 59 also have a "bad weather tour"), and this would decrease the strength of the analysis.
Instead, we prefer to modify the evaluation methodology as presented in answer to comment 3/ and modify the results accordingly.

6/ The performance boost that results in the last column of Table 5 is due to the inappropriate method mentioned above.

As shown by the additional results provided in the answer to comment 3/, taking into account the revision of the methodology proposed by Reviewer#2 leads to results that are similar to those of the present version of the paper for FAR and are slightly lower for POD. Nevertheless, the performance criteria remain satisfactory and confirm the benefit of the IRIP model in identifying sections at risk. Furthermore, considering that mitigation measures only indicate a potential risk and not a proven risk, the final POD obtained by taking Reviewer#2 suggestion into account is the lowest value that can be expected, as it is the most pessimistic way to take into account information about mitigation measures.

In the revised version of the manuscript, we prefer to modify the evaluation methodology as presented in answer to comment 3/ and modify the results accordingly. In particular, we propose to add columns with the intermediate results presented in answer to comment 3/ so that the reader can appreciate the impact of the way mitigation measures are taken into account in the evaluation methodology. The content of answers to comment 3/ will be presented partly in the Results section and partly in the Discussion section.

7/ This aspect is crucial for the entire paper, because the “Results” section concludes with this sentence:

“The results in the last column present very encouraging values, highlighting the added value of the IRIP maps, and of the vulnerability and mitigation measures characterization, for the evaluation of the IRIP model.”

But the last column is affected by this issue, and this casts a shadow on the entire paper.

As shown in the complementary results presented in the answer to comment 3/, a modification of the methodology following Reviewer#2 suggestion does not change dramatically the conclusions of the study and the sentence underlined by Reviewer#2 remains correct, as well as our conclusions that remain supported by the analysis.

Furthermore, we would like to highlight that the present work showed how it was crucial to take into account the information about vulnerability and mitigation measures in the evaluation methodology. In general, only impacts data are considered, because the vulnerability and mitigation measures are more complex to incorporate. In this paper the novelty of the approach is to have included vulnerability and mitigation measures in the methodology. Due to the quality of the data set, we were able to quantify the impact of taking into account or not this information on performance criteria. We propose to add this sentence in the conclusion.

8/ Figure 1 is the same as Figure 4 already published in Lagadec et al., 2018. I understand that you are using the same method. But if a figure has been already published, this should be mentioned in the paper. Moreover, the description of the IRIP method on page 3 uses exactly the same words used in Lagadec et al., 2018. [Lagadec, L.-R., Moulin, L., Braud, I., Chazelle, B., and Breil, P.: A surface runoff mapping method for optimizing risk assessment on railways, Safety Science, 110, 253-267, https://doi.org/10.1016/j.ssci.2018.05.014, 2018.]

Figure 1 is not exactly the same as the one published in Lagadec et al. (2018). As we believe this figure explains clearly the methodology, we prefer to keep it and to mention that the figure is “adapted from Lagadec et al. (2018)” in the figure caption.

In terms of description of the IRIP method, we have already mentioned in the current version of the paper (p.3 line 15-16) that the provided description is mainly borrowed from Lagadec et al. (2018): “The present description is mainly taken from Lagadec et al. (2018) that retained improvements proposed by Lagadec (2017) to the IRIP model.”

We propose to modify the description as follows to make it less similar to that of Lagadec et al. (2018), and to include also the answer to Reviewer#2 comment 9/ and Reviewer#1 comment 3.3/ in the description.

"The IRIP model is briefly described here, but more details can be found in the literature (Dehotin and Breil, 2011; Lagadec et al., 2018). The present description is mainly taken from Lagadec et al. (2018) that retained improvements proposed by Lagadec (2017) to the IRIP model. The IRIP model provides three maps representing three processes involved in storm runoff hazard: generation, transfer and accumulation of runoff. Runoff generation occurs in areas with low infiltration capacity, leading to runoff produced by infiltration excess and/or saturation excess. Runoff transfer occurs in areas where water can be transferred downwards, be accelerated and can produce erosion. Runoff accumulation occurs in areas where water can slow down, concentrate and be accumulated to produce floods and sediment deposits. The IRIP model focuses on runoff occurring outside the river network. It is therefore complementary to flooding risk mapping along river networks. Each IRIP map is produced by combining five indicators derived from geographic information layers (Figure 1, Table 1). Each indicator is classified into two categories: not favorable to runoff, where 0 is attributed to the pixel, or favorable to runoff, where 1 is attributed to the pixel. This yields five binary maps that are then added to create a susceptibility map with 6 levels, from 0 (not susceptible) to 5 (very susceptible). The indicators used for producing each of the three susceptibility maps are presented in Figure 1. The generation map is produced using one indicator derived from a land use map, one indicator derived from the topography, and three indicators derived from a soil map. The indicator related to topography is a combination of the slope and the topographic index (Beven and Kirkby, 1979) and is assigned 1 if both are favorable, and 0 if one is not favorable. The generation map is then considered as an input indicator for the two other maps of susceptibility to transfer and accumulation. This allows accounting for the need of runoff generation, before its possible transfer and/or accumulation. Maps of susceptibility to transfer and accumulation of runoff are produced using mainly indicators based on topography. But the indicators have opposed conditions for being favorable to runoff. For instance, the slope indicator is favorable for transfer in the case of steep slopes, and for accumulation in the case of low slopes. The break of slope indicator is favorable for transfer in the case of convex break of slopes and for accumulation in the case of concave break of slopes. Topographic indicators are computed for each pixel relatively to their upstream sub-catchment allowing to account for upstream to downstream water transfer. The resolution of the susceptibility maps retains the resolution of the Digital Elevation Model (rasterized topography map) used as input data. To determine the thresholds separating the topographic indicator values (slope and topographic index respectively) into values favorable or not to runoff, an automatic classification, the "K-mean clustering method for grids" provided in SAGA GIS was used. The third option that combines two methods: the iterative minimum distance (Forgy, 1965) and the hill-climbing method (Rubin, 1967) to divide the grid values into two classes was retained. The principle of the method is to maximize the inter-class variance, while minimizing the intra-class variance. As the classification is performed using all the grid points located in the study area, the threshold value, separating the two classes (favorable or not to runoff), depends on the study area. The IRIP model can therefore be applied to various territories without a priori local knowledge on the area, as the thresholds can be automatically computed. If local knowledge about threshold values is available, these threshold values can be specified by the user. Note that the choice of the two thresholds for the slope and topographic index has an impact on four indicators out of the 15 presented in Figure 1. Note also that If higher resolution data are included (e.g. Lidar DTM data), it is possible to get information that is more precise and to have explicit representation of linear features such as ditches or roads. In this case, it is not necessary to provide exogenous information about road networks that

are not seen by coarse resolution DTM but are detected with high resolution ones. In this case, adaptation of the model may be required."

9/ Page 3 lines 36-40: Please specify the classification method used. These lines are very unspecific and the conclusions from these lines are not supported.
We propose to modify the text as follows:
"To determine the thresholds separating the topographic indicator values (slope and topographic index respectively) into values favorable or not to runoff, an automatic classification, the "K-mean clustering method for grids" provided in SAGA GIS was used. The third option that combines two methods: the iterative minimum distance (Forgy, 1965) and the hill-climbing method (Rubin, 1967) to divide the grid values into two classes was retained. The principle of the method is to maximize the inter-class variance, while minimizing the intra-class variance. As the classification is performed using all the grid points located in the study area, the threshold value, separating the two classes (favorable or not to runoff), depends on the study area. The IRIP model can therefore be applied to various territories without a priori local knowledge on the area, as the thresholds can be automatically computed. If local knowledge about threshold values is available, these threshold values can be specified by the user. Note that the choice of the two thresholds for the slope and topographic index has an impact on four indicators out of the 15 presented in Figure 1."

10/ Page 5 lines 15-16: this statement is incorrect. The results of the chi-square do not demonstrate that the relationship is highly significant, but that it possible to reject the null hypothesis of independence, because it is unlikely that the null hypothesis of independence is true.
The sentence will be modified as follows:
"A value of χ^2 larger than 10.83 shows that the null hypothesis (independence between the risk levels and the IRIP map) can be rejected at the 0.1% level."

11/ Page 10 lines 2-8: the demonstration (or assumption?) that each section had the chance to experience a rare event is obscure to me.
The point mentioned by Reviewer #2 was raised in the manuscript to support the fact that the chosen case study was adequate to assess the relevance of the proposed evaluation methodology. In particular, the evaluation of the methodology would be biased if the duration of data collection was not long enough so that each section of the railway has had the opportunity to be affected by a heavy rainfall event. In this case, the IRIP model could indicate a risk in a section without reported runoff-related impact because no intense rainfall event would have occurred at that location. To show that the hypothesis that each railway section has had an equal opportunity to experience a high rainfall event, we computed the probability of experiencing [resp. not experiencing] rainfall events of several return periods over a duration of 100 years. This probability is (1-(1-0.1)^100) = 0.99997 [resp. less than 0.0001%] for a 10-year return period, (1-(1-0.05)^100) = 0.994 [resp. less than 1%] for a 20-year return period, and (1-(1-0.0.02)^100) = 0.867 [resp. 13%] for a 50-year return period. Therefore, the working hypothesis is valid and we can conclude that our application of the evaluation methodology is not biased and that the case study was adequate to assess the relevance of the proposed evaluation methodology. We propose to reformulate the sentences as follows:

"Another question is: has each section of the railway had the opportunity to be affected by runoff, i.e has each section of the railway had the opportunity to be affected by an intense rainfall event? If it was not the case, the IRIP model could indicate a risk in a section that would not have been impacted in the absence of any intense rainfall event at that location. To assess the validity of this working hypothesis: "each section had the opportunity to be affected by an intense runoff event", we can

calculate the probability of not having experienced a rainfall event of a given return period during one century. This probability is less than 0.001% for a 10-year return period [$(1/10)^{100}$], less than 1% [$(1/20)^{100}$] for a 20-year return period, and 13% [$(1/50)^{100}$] for a 50-year return period respectively. Therefore, it can be assumed that each section of the railway had the opportunity to experience a rare event at least once during the data collection period. This shows that, if the database is long enough and of course comprehensive (i.e. all the occurred runoff-related impacts were properly reported), the working hypothesis can be accepted and therefore, performance measures can be considered as not biased. In the present case, the comprehensiveness of the database is exceptional, but far from being perfect. However, it was the best that could be collected, and the duration of data collection (more than one century) ensures that the chosen case study was relevant to assess the relevance of the proposed evaluation methodology."

Minor comments:

12/ Page 3 line 39: reduces->reduced Will be corrected.

13/ Page 6 line 15: either…or Will be corrected.

14/ Table 2: please specify also in the table the number of d.o.f for the chi-square test.

The number of degrees of freedom is 1. This will be added to the caption of Table 2.

References

Forgy, E., 1965. Cluster Analysis of multivariate data: efficiency vs. interpretability of classifications, Biometrics, 21, 768-780.

Lagadec, L.-R., Moulin, L., Braud, I., Chazelle, B., and Breil, P.: A surface runoff mapping method for optimizing risk assessment on railways, Safety Science, 110, 253-267, 2018.

Rubin, J., 1967. Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem, J. Theoretical Biology, 15,103-144.