Reviewer 1 – Giuseppe Esposito

The manuscript is a solid and interesting contribution documenting post-fire debris flows in New Mexico (USA), where the research team is dedicating considerable efforts for understanding both predisposing and triggering factors that control the initiation of these processes, as reported in their recent papers and datasets. The current work presents a series of results deriving from both monitoring and modeling activities which are well depicted and supported by a robust bibliographic background. In particular, original measurements collected in the field make the research outcomes noteworthy and helpful for other scientists that are investigating the same hazard in the USA and worldwide. The overall quality of the manuscript is very good. I have just a few questions and suggestions reported in the attached pdf file that may contribute to improve the article before considering it for publication in NHESS.

In summary, the Authors carried out both monitoring and modeling activities but the related description is often mixed in a way to generate some confusion in the reader. This is evident in the description of objectives and methodologies. Minor corrections to the text and additional photographs could clarify some steps which are described too succinctly in the methodology section.

The Authors provided a sufficient number of Figures of good quality. However, I suggest to improve the Figures 1 and 3 by adding some relevant contents as highlighted in the annotated pdf file. Further tables or charts could be added in the results section to better display characteristics of triggering rainstorms.

In the discussion section, I have noted that the Authors refer mostly to the USA-related literature. I suggest to consider also valuable studies performed in other continents, since several findings (e.g., concentration of PFDFs in the first months after fires) show relevant similarities with other settings. This may emphasize the article impact in the future. More comments about lower runoff rates in areas burned at moderate/high severity in the first months after the fire are welcome.

R: Thank you for your review of our manuscript. Below, respond to each comment made in the pdf. Comments are identified by their corresponding line number, table number, or figure number in the pdf.

Line 33: Clarify which type of response.

R: Changed to “As a consequence, burned watersheds are more susceptible to debris flows.”

Line 34: Months or days too.

R: Yes, thank you for emphasizing this point. We have added the following phrase to highlight that there may be very little time between a fire and the initiation of PFDFs: “…, sometimes even prior to the fire being contained (e.g. Kean et al., 2019)…”

Line 40: This part seems too long and confusing. Rephrase please.

R: We have rephrased and broken the original sentence into two separate sentences: "PFDF hazards, which are already well recognized in the southwestern USA (Staley et al., 2020), are likely to increase in the future due to increases in both area burned (Holden et al., 2018; Singleton et al., 2019) and the frequency of extreme precipitation (Kirchmeier-Young and Zhang, 2020). Identifying when and where debris flows are most likely to initiate within burned areas will help to better assess hazards and prioritize mitigation efforts."

Line 46: The connection between watersheds susceptible to debris flows and flow behaviour is not clear. Clarify please.

R: We inserted an additional sentence to clarify: "It is important to identify watersheds that are susceptible to debris flows since the high sediment concentrations and peak discharges associated with debris flows may require additional or different strategies relative to those used to mitigate negative effects of flood flows. The high sediment concentration in debris flows changes their flow behaviour, resulting in coarse-grained flow fronts with peak discharges and flow depths that can exceed those expected from water-dominated flows (Kean et al., 2016)."

Line 81: I suggest to present also the case of low severity wildfires that you have well addressed in this paper: https://doi.org/10.1029/2020JF005997

R: Reference added.

Line 125: Since I found that you mixed both monitoring and modeling activities, I believe that a distinction between monitoring- and model-based objectives is necessary here.

R: To meet our objectives, we do use a variety of methodologies, including modelling, field monitoring, and in-situ measurements. We hesitate to break out separate objectives based on methodology since we think that it is the integration of data from these different sources that allows us to gain insights into postfire debris flow processes.

Line 127: Specify how many months/years

R: Changed to "…the first three monsoon seasons (approximately 2.5 years) following the fire…"

Line 130: Specify whether you want to reach this goal in the study area or elsewhere please

R: Rephrased to "An overarching goal of this work is to provide data and process insights to improve situational awareness of PFDF hazards, particularly in the southwestern USA. More broadly, data collected as part of this study adds to a growing set of PFDF observations from around the world that can inform data-driven models designed to assess the potential for PFDFs."

Line 136: Information on the soil type affected by erosion processes should be inserted.

R: We added the following: "Dominant soil types include Mollisols, Inceptisols, and Alfisols and the soil texture is classified as a loam (43% sand, 45% silt, 12% clay)."

Line 143: Add in the caption some information about the soil burn severity in the background

R: We added the following: "Soil burn severity (SBS) varied throughout the study area. SBS was assessed by the Burned Area Emergency Response team using methods that rely on both field observations and the difference normalized burn ratio."

Line 155: incised during

R: Done.

Line 155: rainstorms of

R: Done.

Line 161: A tabel summarizing the used thresholds may be useful for the reader

R: We do not know the specific dNBR thresholds used to generate the soil burn severity map. These thresholds vary from fire to fire and are determined by the Burned Area Emergency Response team.

Line 165: Distinguish between monitoring and modeling. I suggest to avoid to mix both activities in the same sections

R: We now separate the methods section into two main subsections, Sect 3.1 and Sect 3.2. Sect 3.1 has several sub-sections and focuses on methods related to flow monitoring and field measurements. Sect 3.2 has two sub-sections and focuses on methods related to modeling debris flow likelihood and runoff generation.

Line 173: I suggest to indicate some altitudes in Figure 3 or to classify the DTM in background based on elevation data

R: We have added elevation contours to Figure 3.

Line 177: This choice needs to be motivated

R: The decision to focus on monitoring only the five intensively monitored watersheds in subsequent years was driven by time and resource constraints. To clarify, we have added: "However, due to time and resource constraints…."

Line 182: Insert here a brief sentence introducing the following sections

R: We have added the following: "In Sect 3.1, we describe methodologies related to field measurements and flow monitoring, including estimating ground cover and infiltration capacity, monitoring rainfall and flow activity, and analyzing rainfall characteristics. In Sect 3.2, we describe modeling methodologies used to assess debris flow likelihood and temporal variations in runoff generation as a function of time since fire."

Line 201: I do not understand where these measurements were performed. You mention both "varied locations" and in correspondence of the vegetation transects. Clarify please

R: Measurements were made opportunistically in 2020 at arbitrary locations within the study area. We made some measurements in areas burned at low severity and some in areas burned at moderate/high severity. We have rephrased this sentence for clarity: "Measurements were performed at arbitrary locations…"

Line 204: Does each group refer to each meter of the transect?

R: By "group," we are referring to a set of measurements made in an area with a given burn severity at a particular time (e.g. all measurements made in an area burned at low severity in May 2021). We added the following to clarify: "…(i.e. in an area burned at a given severity at a particular time)…."

Line 263: Some explanations about this choice could be helpful

R: The point scale modelling allows us to quantify the combined effects of changes in K_{fs} and h_f to runoff generation. A 2d model that involves routing water across the landscape is not essential for this type of analysis and would add substantially to computation time. We have added the following sentence to clarify: "While point scale modelling does not allow us to assess the concentration of runoff across the landscape, it does allow us to assess the combined effects of changes in K_{fs} and h_f on runoff generation in response to different rainfall intensities."

We have also added the following as motivation for the choice of the Green-Ampt infiltration model: "The Green-Ampt model represents infiltration-excess overland flow, which is the primary runoff-generation mechanism during storms that produce runoff-generated PFDFs in the southwest USA (Gorr et al., 2023; Schmidt et al., 2011). The model has been widely applied to simulate postfire infiltration and runoff generation (Van Eck et al., 2016; Ebel, 2020)."

Line 272: This has been already used before in the sentence. Consider to substitute with "after"

R: Changed to "…time after fire…"

Line 289: Add some information about data communication systems, if any, or if you downloaded the datasets on site

R: We added the following to the end of the first paragraph in this section: "Data were not telemetered but were periodically downloaded on site."

Line 299: These are not indicated in Figure 1. Revise please

R: We have added camera locations to Figure 1.

Line 300: If one considers the time receorded in the photos only, the debris flow time can be thus affected by a 3-min uncertainty? If so, underlain this.

R: We agree that using the camera to estimate debris flow timing within rainstorms could result in additional uncertainty relative to using the data from pressure transducers. We did not use information from cameras to estimate the timing of debris flows within rainstorms.

Line 330: Is this the total rainfall recorded between the debris flow observation time and the storm initiation? Clarify please

R: We added the following sentence to clarify: "We computed I_D at intervals of 1 minute throughout each rainstorm."

Line 337: This seems quite tricky. In the first sentence of the chapter, you mentioned durations ranging from 5 to 60 minutes but here you refer to 15-minute duration on the base of previous studies. Clarify please

R: We do compute rainfall intensity averaged over durations ranging from 5 minutes to 60 minutes. We focus on reporting results using rainfall intensity averaged over a 15-minute time period since this duration has been identified by past studies as being particularly relevant for predicting debris flow initiation in small, burned watersheds. A separate issue is one of how to choose the triggering rainfall intensity. In line 337 of the original submission, we state that we compute the triggering intensity by looking at the peak rainfall intensity within a 15-minute window prior to debris flow initiation. As an example, the triggering 5-minute intensity would be found by determining the peak in I5 between t* and t*-15 minutes, where t* denotes the time the debris flow passes over the pressure transducer. We have added this example to the manuscript to clarify: "For example, the triggering I5 would be equal to the maximum value of I5 between t* and t*-15 where t* denotes the number of minutes following the start of the rainstorm when the debris flow was detected at the watershed outlet."

Line 391: I disagree with this approach, since you may introduce a strong bias in the analysis

R: This comment is in reference to the following sentence "In cases where we determined that a debris flow occurred but we could not constrain the timing of debris flow, we assigned the triggering intensity to be equal to the peak rainfall intensity observed in any storm prior to the debris flow survey."

Assuming that the triggering intensity can be approximated by the peak rainfall intensity is not ideal, but it is standard practice when defining rainfall ID thresholds in cases where there is no way to constrain the timing of the debris flow within a particular rainstorm. As the reviewer correctly points out, past studies have demonstrated that this assumption can bias rainfall intensity-duration thresholds to be high which could lead to underestimation of debris flow hazards when these thresholds are applied in the future (Staley et al., 2013; Raymond et al., 2020). We have added the following to clarify our approach: "Approximating the triggering intensity, which must be equal to or less than the peak rainfall intensity, by the peak rainfall intensity can lead to overestimation of rainfall-intensity duration thresholds for debris flow initiation (Raymond et al., 2020). In cases where we were able to constrain the timing of debris flows within rainstorms, we examine differences between the triggering and peak rainfall intensities."

In the nine cases where we were able to constrain the timing of debris flows within a rainstorm, we provide results that quantify the differences between triggering and peak rainfall intensities. For example:

"In the nine cases where we were able to determine debris flow timing within rainstorms, we computed the triggering I15 and found that it ranged from 33-76 mm/h (Table 3). In four of the nine cases, the peak and triggering I15 were the same (Table S2). In the five remaining cases, the difference between the peak and triggering I15 was 43, 38, 1, 2, and 10 mm/h (Table S2). Storm cumulative rainfall totals were also greater than storm rainfall totals prior to debris flows, with the most substantial difference (31 mm) occurring during the storm on 9 September 2020 (Table S2). On average, the debris flow triggering time (i.e., the time the debris flow was observed at the outlet) was approximately 3 minutes before the time of the peak I15 (Figure 7). The debris flow triggering time preceded the peak I15 in six out of nine instances. Debris flows passed the watershed outlet, on average, less than 1 minute following the time of peak I10. In contrast, debris flow triggering times preceded the time of peak I30 and I60 by roughly 13 and 31 minutes."

"The recurrence interval of peak 15-minute rainfall intensities during debris flow-producing storms ranged from 0.5-7.5 years with a mean of 3.4 years. In contrast, the recurrence interval of 15-minute rainfall intensities that triggered debris flows (i.e., only including observations where we have flow timing data) ranged from 0.5-3.4 years with a mean of 1.3 years (Table 3)."

In summary, we agree with the reviewer that this assumption can bias results and we were able to quantify the extent to which that occurred at our site. This provides additional guidance for interpreting rainfall ID thresholds in this region that have been developed using data that does not allow for identification of debris flow timing within rainstorms.

Line 398: It could be interesting to include a photograph of deposits to highlight how did you perform the pebble counts

R: Unfortunately, we don't have any field photos taken during our pebble counts at the Tadpole Fire. We do have photos from other sites where we have applied the same methodology and we have added one of these as Figure S2.

Line 406: This should be also reported into the chapter 2 where you describe the study watersheds

R: We decided not to include this information in section 2, where we provide a general description of the study area, because at that point in the text we have not yet described the installation of the monitoring equipment.

Line 413: How did you get KF factor for the study area?

R: The KF factor comes from STATSGO database. We have added the following for clarity shortly after the KF factor is introduced: "…from the STATSGO database (Schwarz and Alexander, 1995)…"

Table 1, column 6, row 2: 95?

R: Thank you for pointing this out. The values listed in Table 1 are correct as reported, but we realize that our methods for determining total ground cover were not clearly described. We have added the following to the methods section just prior to our original definition of total ground cover ("The percentage of total ground cover was determined based on the number of first hits that were classified as either canopy or litter while bare ground consisted of all measurements where the first hit was soil or rock."):

"At each measurement location (i.e. every 20 cm along the transect), we recorded whether there was understory canopy cover, litter, soil, or rock. It is possible for a measurement to indicate the presence of understory canopy, litter, and either soil or rock. In other words, if understory canopy was present, we still assessed the presence or absence of litter, soil, and rock underneath the canopy. If both canopy and litter were present, we continued to determine the presence of either soil or rock."

As a result, the total percent ground cover is not determined by summing the percent litter cover and percent canopy cover since litter and canopy cover may co-occur in the same location. In row 2 of table 1, the percent canopy cover is 94%, litter cover is 1%, and total ground cover is 94%. This implies that the 1% of cases with litter overlapped with areas of canopy cover. Therefore, total ground cover is still 94%. As we explain in the revised text, "We adopted this definition of total ground cover since it reflects the percentage of the ground surface that would be exposed to direct raindrop impact. Exposure of bare ground may affect processes such as raindrop-induced sediment transport and surface soil sealing that are influential in recently burned areas (Larsen et al., 2009)."

Table 1, columns 4 and 5, row 6: Revise please!

R: See the response to the above comment. We have added text to clarify our methods for determining total ground cover.

Line 459: Revise please!

R: We rephrased some text and broke this single sentence into two sentences. Changed to "At the surface of soils burned at moderate or high severity, approximately 55% of WDPTs indicated moderate or extreme water repellency. In soil burned at low severity, 33% of measurements indicated moderate or extreme water repellency."

Line 467: This is surprising. If you have some interpretation on this add one or more sentences in the discussion section.

R: This comment refers to the following sentence: "The point-scale rainfall-runoff model constrained by the minidisk measurements indicates that runoff ratios in areas burned at moderate/high severity were lower or similar to those simulated under unburned soil conditions after 0, 10, and 26 months of recovery." We have added the following to the discussion:

"In an analysis of data from southern California, USA, Ebel and Moody (2020) found that the ratios of K_{fs}, S, and h_f in burned to unburned soils were 0.37, 0.36, and 0.66. Substantial variability exists from site to site (Ebel, 2019), however, with postfire K_{fs} sometimes being greater relative to that in nearby unburned soils (Raymond et al., 2020). Collecting additional information related to fire effects on soil physical and chemical properties could help explain variability in how soil infiltration capacity changes in response to burning (Ebel et al., 2022), though this was beyond the scope of our study."

Line 474: Indicate in how many watersheds.

R: Done. "We observed 16 debris flows from 11 different watersheds…"

Line 496: This is new for me. I did not read on these samples in the methodology section. Please add the related information.

R: Collection of sediment samples from debris flow deposits is described in the last paragraph of Section 3.5, but we did not mention collection of the hillslope samples in the original submission. We have added the following sentence to Section 3.5 to complete the description of our sample collection: "For comparison, we also collected samples from the upper 5 cm of mineral soil from two burned hillslope locations."

Line 535: Indicate in which ones

R: This information is contained in Table S2 so we have added a reference to Table S2 at the end of the sentence.

Line 536: Which is this table?

R: This is referring to Table S2 in the supporting information. This table contains a summary of debris flow timing information and triggering rainfall characteristics, including triggering 15- and 30-minute rainfall intensities, peak 15- and 30-minute rainfall intensities, cumulative rainfall prior to debris flow initiation, and storm cumulative rainfall.

Line 539: Can this statement be extended to all of the nine cases? If so, you should mention that the triggering I15 preceded the peak I15 in all of the 9 cases.

R: The triggering I15 did not always precede the peak I15. We have now clarified this with the statement: "The debris flow triggering time preceded the peak I15 in six out of nine instances."

Line 540: Is this a new metric in the article? The peak I10 occurred before or after the debris flow triggering time?

R: This is the first time that we have mentioned I10, though we introduced the more general notation of I_D earlier in the manuscript (equation 10).

Line 540: It is difficult to imagine this without a plot

R: It is possible to visualize the relationship between the timing of the peak I15 and the timing of the debris flow at the watershed outlet by looking at Figure 7. In the main text, we have added a reference to Figure 7 at the end of the sentence.

Line 585: Similar outcomes were found also in other parts of the world, like in Europe. Please add some reference on studies documenting PFDFs in the first months after wildfires outside the USA.

R: We have added the following sentence and 5 references: "Observations from around the world similarly indicate that runoff-generated PFDFs tend to occur primarily, though not exclusively, in the first year following fire (Wang et al., 2022; Jin et al., 2022; Esposito et al., 2023; García-Ruiz et al., 2013; Jordan, 2016)."

Reviewer 2 – Don Lindsay

The authors did a great job illustrating the importance of ground cover over infiltration capacity in triggering PFDFs. I believe the study approach they applied will provide a template and foundation for other scientists in the field to emulate and build from, respectively. I appreciate the concise, easily-digestible writing style they applied. In addition, I thought the figures were well thought out and do a great job presenting the data. I provide minimal comments in the attached pdf for the authors to consider.

Line 426: I recognize the USGS suggests using P=0.5 to assess PFDF hazards under emergency conditions; however, because you are ultimately comparing the M1 model results to actual debris flow results, I believe it would be good to also show measured triggering I15 against the M1 model results with P=1.

R: Thank you for this suggestion. As the reviewer notes, it is common to set p=0.5 when using a logistic regression model as a classifier (i.e. to determine whether or not a debris flow will initiate). Staley et al. (2017) set p=0.5 when using the M1 model to determine a rainfall ID threshold in their study. Setting p=1 would result in an undefined rainfall ID threshold, but we can compare the observed debris flow triggering intensities with the rainfall intensity required to achieve p=0.9. Defining an ID threshold based on when the M1 model returns a likelihood of 0.9 would be considered very conservative. Below, we summarize a few results focusing on rainfall thresholds computed for a 15-minute duration using p=0.5 and =0.9.

Defining the threshold based on when p=0.5 and p=0.9 results in mean M1 modeled thresholds of 18 mm/h and 30 mm/h, respectively, for the watersheds where we can constrain debris flow timing. On average, the M1 modeled threshold defined by p=0.5 is 34 mm/h less than the observed triggering intensity for the 9 debris flows at our site where we can constrain the debris flow timing within storms. Similarly, the M1 modeled threshold defined by p=0.9 is 23 mm/h less, on average, than the observed triggering intensity for the 9 debris flows at our site where we can constrain the debris flow timing within storms. This supports our conclusion that the M1 model underestimates debris flow triggering intensities at our site. Since determining debris flow initiation thresholds using p=0.5 and p=0.9 does

not change the relative susceptibility ranking of watersheds (i.e. which watersheds have lower or high M1 thresholds remains unchanged regardless of the choice of p), these calculations also do not change our conclusion that the M1 model does well at assessing the relative susceptibility of different watersheds to debris flow responses. Given the consistency of results regardless of the choice of p, we have decided not to make a change to the manuscript and only presents results using the more standard value of p=0.5.

Line 498: It would be good to identify the range of grain sizes in the pebble count and the sieve analyses. At one point I believe the point count was performed on grain sizes >2mm. However, it appears here that the point count was done on grain sizes >20mm. Some clarification would be good.

R: The D50 referenced in line 498 was determined from sieve analyses of debris flow sediment. We used sieve sizes of 32 mm, 16 mm, 8 mm, 4 mm, and 2 mm. The D50 ranged from < 2 mm to roughly 20 mm. We have added information about the sieve sizes to the methods section: "These samples were air dried and sieved, using sieve sizes of 32 mm, 16 mm, 8 mm, 4 mm, and 2 mm…"

Line 570: "The temporal distribution of rainfall within rainstorms that produced debris flows (red lines) are similar, with the majority of rainfall occurring during the second quarter of the storm duration (0.25 ≤ normalized time ≤ 0.5)." This is interesting. This implies that there may be a linkage between initial abstractions before runoff occurs that then trigger PFDFs. This is partially suggested by the equally steep curves left of the red curves that have similar peak depths. The only apparent difference between the two sets of curves is the presence of low rainfall in Q1 for the red curves to infill voids prior to initiating overland flow. Moody, 2012, covers this a little by assuming some initial rainfall goes towards saturating surface depressions and near surface voids before runoff develops. This may be a compelling observation that may warrant additional consideration for future studies dedicated to triggering rainfall events.

R: Thank you for this comment and the opportunity to clarify some aspects of the SRP results. When interpreting results of the SRP analyses, it is important to keep in mind that only the normalized rainfall depth is used. We have added a clarifying statement to the caption of Figure 9: "Note that the standardized rainfall profiles are plotted using normalized rainfall depth so the curves do not provide information of the absolute value of rainfall depth during different portions of the rainstorm."

Since the SRP analyses do not depend on the absolute value of rainfall, it is challenging to use them to assess the role of initial rainfall abstractions in the debris flow generation process. There are grey curves to the left of the red lines in Figure 9 that are equally steep, but this only provides information about the temporal distribution of rain within the storm and is not indicative of the overall depth of rainfall or the depth of rainfall that occurred early in the storm. The SRPs are helpful for a quick visual assessment of the type of rainfall event (e.g. convective vs frontal system) that did or did not trigger a debris flow. This technique was recently applied by Esposito et al. (2023) in their study of PFDFs in Italy. We think that analysis of SRPs is one method that may help assess similarities and differences among the types of rainfall events that trigger PFDFs in different regions. We added the following to the discussion section:

“Standardized rainfall profiles of debris-flow producing storms generally plotted above the 1-1 line (black line in Figure 9b), which is a characteristic associated with convective rainstorms (Esposito et al., 2023). This finding is consistent with the timing of debris flows during the summer months shortly following the fire when convective rainstorms associated with the North American monsoon are common in the region. Esposito et al. (2023) similarly found that storms that produced PFDFs in Italy had SRPs consistent with convective rainstorms rather than frontal systems.”