Reply to the interactive comment by Referee #3

We thank the anonymous referee for her/his comments, which helped to clarify a number of points and improve the revised version of the paper. The reviewer's comments are quoted above the author's responses and are marked in italic.

General Comments:
This manuscript examines slope stability predictions from the Borga et al. model at different mapping resolutions, and compares them to observed landslide distributions in a 2.7 square kilometer Sicilian catchment that was hit by intense storms in 2007 and 2009. The model performance is shown to be better when the source DEM has a resolution of 4m or 10m than when it has a resolution of either 2m or 20m. It is good to see this work, because natural hazard models are not tested against real-world data as often as they should be, leaving both scientists and practitioners uncertain of what their strengths and weaknesses are. In part this reflects the scarcity of good comparison data sets, particularly for hazards that are episodic and rare. It may also reflect an implicit judgment among researchers that the academic career "payoff" will be greater from writing yet another model than from testing existing models.

Specific Comments:
The abstract highlights two main results: (1) Model performance is better when the source DEM has a resolution of 4m or 10m than when it has a resolution of either 2m or 20m. (2) Model performance is worse for road-related failures (the model does not specifically represent the effects of roads). The rhetorical extrapolations from these results (the last two sentences of the abstract), however, are not supported by the results themselves, as far as I can see. I consider these two sentences in turn.
(1) "These findings show that shallow landslide predictive power can benefit from increasing DTM resolution ONLY WHEN the model is able to describe the physical processes emerging at the smaller spatial scales resolved by the digital topography" (emphasis added). This may in fact be true, but I cannot see where this is demonstrated in the paper. No attempt has been made to determine which physical processes emerge at which spatial scales. If this could be done, and if those processes could be captured in a model, and if this resulted in better model performance at higher DTM resolution – all of which would need to be demonstrated, not just asserted – then one would be justified in making the statement above... but as "when", not as "only when". To make the "only when" assertion, one would need to additionally demonstrate that there is no other way to improve model performance at higher DTM resolution. Nothing remotely close to this has been done. The authors seem to have overlooked a very simple explanation for the lower performance at higher DTM resolution (at least, I didn't see it): this is an infinite-slope model, and therefore it assumes that the soil surface and the lower boundary are parallel to one another. This must be approximately true when slope is averaged over distances that are long compared to the soil thickness (which at this site is reportedly 0.7-3m). But over distances that are not long compared to the soil thickness, the infinite-slope model is simply not correct. This isn't a case of a new "physical process emerging at smaller spatial scales" (in the language of the abstract), it's a case of a model approximation that's not valid at smaller spatial scales.
We agree with the Reviewer and revised the Abstract and manuscript accordingly.

Another obvious explanation for the poorer performance at smaller spatial scales is that edge effects (specifically, lateral root cohesion at the edges of a potential slide) have been ignored in this model. Again, this obvious explanation is not discussed in the paper. And again, this is not a "physical process emerging at smaller spatial scales", but rather a model approximation that

becomes worse as the spatial scale decreases. The argument that this is done (p. 9767) "to avoid making assumptions about landslide size" is not technically correct. Correctly accounting for lateral cohesion will provide information about landslide size as a model output (which is rather the opposite of making an assumption about it, as a model input). This is not to say that doing so would be easy, of course.

We agree with the Reviewer and included the mentioned model approximation among the reasons for the results we found.

The second rhetorical extrapolation is (2), "Model results show that the combined use of high DTM resolution and a model capable to deal with road-related processes may lead to substantially better performances in landscapes where forest roads are a significant factor of slope stability." This statement is only true if "may" means "may or may not"... in which case it is also trivial and meaningless. In any case, the model results only show that a model that does not incorporate road effects does not predict landslides were roads are an important causal factor. This is not at all surprising, but it also does not demonstrate that another model would perform better. That may be true, but it is simply incorrect to say that the current "model results show" this, so the conclusion is simply unsupported by the paper.

We agree with the reviewer and revised the abstract accordingly.

Reading the abstract, the reader would be unaware of perhaps the most dramatic result of all, namely that the model greatly overestimates the risk of landsliding. Between 6 and 14 percent of the landscape is classified as unconditionally unstable (and thus should fail even without any rain), and an additional 14-18 percent of the landscape is predicted to fail in modest storms (return interval <10 years), but only 2.7 percent of the landscape actually did fail, in response to massive storms with return intervals of 100-300 years. This certainly needs to be highlighted in the abstract, because otherwise it looks like the authors are trying to hide an "inconvenient truth".

We revised the abstract reporting this result.

The paper presents the percentages of landslide area, and of total catchment area, in different stability categories, but those are not the percentages that matter from a risk-assessment standpoint. What matters from a risk-assessment standpoint is the fraction of the landscape in each stability class that actually did fail, and the fraction that did not. It is almost impossible to figure this out from the data presented in the tables, which seem to be organized so that they are deliberately obscure in this regard. This really needs to be fixed. The many tiny tables, each presenting just a fragment of the overall picture, should be replaced with two comprehensive tables: one that presents the percentages of the whole landscape, and of the landslide areas, that fall in each risk class, and a second table that presents the percentage of the area in each risk class that actually failed. This is likely to paint a rather different picture than the one the reader gets now.

We revised the Tables, providing the information requested.

One can estimate that of the 6 to 14 percent of the landscape that were classified as unconditionally unstable, only about 1 to 1.5 percent actually did fail (in a very extreme storm). Of the remaining 14-18 percent that were classified as prone to failure in modest storms, only about another 1 percent or so actually failed. So the "false alarm" rate here is rather high (particularly considering that if the model were accurate, all of these points should have failed long before the study even began)! The explanation for this overprediction that the authors provide on page 9778 is unconvincing. Of course the input data are uncertain, but the results show strong bias rather than random noise. Also, the authors' statement that "we consider this prediction as an indication

for likely failing sites in future storms rather than areas proved stable during previous storms"
rather clashes with the data showing that if the model predictions were accurate, these sites would
have failed long ago.

We agree with the Reviewer that our statement that "...this prediction as an indication for likely failing sites in future storms rather than areas proved stable during previous storms" requires a more thorough explanation, as follows.

Accordingly with the Reviewer, two main types of error can be identified by comparing observed landslides and areas characterised as unconditionally unstable or stable by the model: 1) a site is characterized by the model as unconditionally unstable or with a low value of critical rainfall, but no scars were observed on it; 2) a site is predicted as unconditionally stable or characterised by a high value of critical rainfall, but slope failures were mapped on it (see also Casadei et al., 2003). The first type of error indicates that the model over-represents areas potentially subject to shallow landsliding. Such error typically arises from the action of processes external to the model framework and local conditions specific to the study catchment. Reasons for this mismatch could include errors in the critical rainfall calculation arising from the assumption of a given soil depth exceeding the actual one. However, it should be observed also that type 1 error can result from concealing of areas failed at some time in the past. For instance, our observations in experimental basins (Tarolli et al., 2011) indicate that small, shallow debris flow scars rapidly heal and are difficult to detect after as few as three-four years. These findings suggest that unfailed slopes characterised by the model as unconditionally unstable or with low critical rainfall should be interpreted as likely sites of failure in the future rather than areas presumed to be tested and proved stable during previous storms.

Casadei, M., Dietrich, W. E., and Miller, N. L., 2003. Testing a model for predicting the timing and location of shallow landslide initiation in soil-mantled landscapes, Earth Surf. Processes Landf., 28, 925–950.

Another conspicuous issue that the manuscript avoids discussing is that the study focuses on a very small area (2.6 square km) where many landslides actually occurred. But what about the surrounding landscape? The reader is told that the study area was 6 km from the two rain gauges; this means that there is an area of about 200 square km that is at least as close to those rain gauges as the tiny area studied here. Surely there is topographic information for this larger area as well, and surely there are aerial photographs that would allow one to tell where landslides occurred. There may be good reasons why this particular tiny area was studied, but frankly the reader is left with the impression that the authors have a model that vastly overpredicts landslide risk, and therefore deliberately selected a very small place where a very large number of landslides actually occurred. This is troubling, because it raises the possibility that the "false alarm" rate in the surrounding terrain could be much higher than implied even by the results presented here.

We agree with the Reviewer 1 that the choice of the study area deserves a more thorough explanation. We selected this study area for a number of specific reasons: 1) the area was impacted by the extreme event in 2010, which caused 36 casualties; hence more resources were made available on this area for field data collection; 2) the area presents a good combination of road network density and road-related landslides; 3) larger scale analysis of road related geometric features and landslides was unfeasible with the project resources.

Technical comments:
The equation between 6 and 7 (why isn't it numbered?) is obviously wrong.
Done. In the revised version we fixed this issue.

T is transmissivity; presumably the authors mean T_R (the recurrence interval)? Note also that (assuming that T really should be T_R) the math blows up when T_R is less than 1. Assuming T_R is a dimensional quantity (years, for example), this makes no sense (just measure time in centuries or millennia, and it is very easy indeed for T_R to be less than 1. What is the justification for this ln(ln()) formulation (and the corresponding exp(exp()) formulation in equation 8? In real-world systems it is very rarely seen.

Eq. 7 has been corrected for the typo (T is the return time Tr, indeed). We thank the Reviewer for reporting this error. We have adjusted the equation number as well.

The double exponential model for extreme analysis dates back to the model introduced by Gumbel (1954). The model has been used successfully in many fields, from floods to weather variables such as rainfall, temperature and atmospheric pressures, from radioactivity to stock markets (Kottegoda and Rosso, 1997).

Gumbel, E.J., 1954. Statistical theory of extreme values and some practical applications. Applied Mathematics Series, 33. U.S. Department of Commerce, National Bureau of Standards.

Kottegoda, N.T., Rosso, R., 1997: Statistics, Probability, and Reliability for Civil and Environmental Engineers. McGraw-Hill Companies, 735 pp.

On page 9772 it is argued that the high length/depth ratio of the landslides which actually occurred "ensures proper application of the infinite length assumption within the infinite slope stability model". Sorry, this is just illogical. The fact that the actual landslides are long and thin does not validate the model's assumption that all *potential* landslides are also long and thin... for the simple reason that this assumption creates many potential landslides under conditions where actual landslides do not occur (for example, short, steep slope segments that will fail in the infinite-slope model will *not* fail in the real world... precisely because they do not go on for infinite length!). Natural hazards models need to evaluate all *potential* failures and correctly predict which of these are likely to become actual failures. Many of the potential failures will not become actual ones because they lack one or another characteristic of the actual failures (for example, they are not long enough or wide enough). Therefore it makes no sense to assume that all the potential failures have all the characteristics of actual failures. This will lead to a systematic overestimate in the risk of failure.

Consider the following hypothetical example. Let's assume that the real-world rule that Mother Nature follows is that any slope steeper than X% and longer than 200m will fail. That means that all observed failures will be longer than 200m. But this does not validate the infinite-slope assumption, which holds that all slopes are infinite and therefore any slope steeper than X% will fail. In fact, many slopes will be steeper than X% but shorter than 200m, and thus will not fail.

This sentence has been revised and rephrased. Here the reviewer likely misinterpreted our intent. Milledge et al. (2012) provided a very detailed overview about the suitability and limits of infinite length assumption and grid cell size. Here we simply report that the characteristics of the examined landslides are suitable with the infinite-slope length assumption, according with the findings of Milledge et al. (2012).

-9762-26 evolve INTO debris flows
Done.

-9763-4 landslide (no s) catalogues.
Done.

- 9763-16 steady-state and infinite-slope (need hyphens)
Done.

- 9765-2 of forest ROADS
Done.

9765-8 "may impact on the predictive power" – how about "may AFFECT the predictive power"? "impact on" means "crash into".
Changed.

9766-2 intersected, not interested. but the word you really want is probably something else, perhaps "dissected".
We changed it into 'characterized'.

9766-17 "contributes subsurface flow" (no "as")
Done.

9767-2 impeding, not impending
Changed.

9767-18 missing "and"
Added.

9769-2 what is GEV?
We included the definition of the acronym ("Generalised Extreme Value") in the Section title.

9770-23 road system HAS a length of ## and a density of ##
Changed.

9770-26 drains through CULVERTS into THE valley bottom. The road width RANGES between...
Done.

9772-7 causing several fatalities (not victims).
Corrected.

9772-10-11 grammar error
Corrected.

9772-21 what's vegetation "surcharge"? (the word means "extra payment")
The term surcharge is the correct technical term to mean the tree weight per unit area to be included in the infinite slope equation.

9775-28 confusing word order
Changed.

9778-15 "2m and 2m"?????
The firs one is 20 m. We corrected this error.

9779-5 "the effects of external processes may be dependent themselves on the resolution of the digital topography" What? The external processes (in the real world) do not know or care about the scale of our digital maps!

The sentence has been rephrased.