

Some ideas for emulators and sampling rare events in the pyroclastic flow problem — a working document

SAMSI granular flow group
Elaine Spiller

March 26, 2007

Our ultimate goal is to draw a probability map around a volcano where contours represent the probability of a flow over some critical value, say 2m, over a substantial length of time, say 100 years. We are specifically interested in mapping regions of very low $P(h_{flow} > h_{crit})$ to determine safe locations for municipal development (e.g. hospitals, schools). Our main tool in this mapping is the TITAN software which is a computer model of pyroclastic flows.

This problem raises some significant statistical challenges

1. We seek to predict the probability of events for which we hope to never have data, i.e. catastrophic events.
2. Data collection for large volume flows is a bit spotty due to obvious safety concerns, and data for frequent yet small volume flows is “pre-binned”. However, Robert and Simon have been making some progress on characterizing a distribution for the flow volume.
3. The TITAN software is moderately expensive (~ 1 hour per sample). This rules out any hope of using TITAN output directly as samples in a Monte-Carlo (MC) simulation of rare events, even if utilizing variance reduction techniques (importance sampling, etc).
4. There are several issues in designing emulators of the TITAN model for this system.
 - a) Zeros – the output from one run will typically have no flow over a significant portion of the land area we wish to map.
 - b) Designing a “spatial” emulator, i.e., one that can interpolate output of interest (max height, say) at untested locations around the volcano is challenging both due to the size of the design problem and the zeros mentioned previously.

1 Proposal for a “simple” emulator/sampling problem: Calculate $P(h_{max} > h_{crit})$ at a single spatial location of interest

First lets consider the model inputs:

1. Initial mass location (stochastic) — East and North coordinates, each uniformly distributed
2. Flow volume (stochastic) — Volume of flow, Pareto distributed with parameters determined from flow data by Robert and Simon.
3. Initial flow velocity (stochastic?) — Starting velocity of flow. Currently taken to be zero.
4. Friction — two sorts (?) one material dependent – friction between interacting grains of material, one friction of material interacting with the ground – depends on rainfall.
5. any others?

More thoughts on model inputs (specifically 1-3):

1. Currently at Montserrat there is a “pimple” forming and when it breaks it will (presumably) be a large volume event. Q – Is the location of such pimples typically random or are there specific spots where they tend form? If the location is random, perhaps the initial mass location is the correct input to vary (if we are looking for probabilities over a fairly long time period, say ~ 100 years). In either case we need to be a bit more careful, but if pimple locations are not random perhaps we need to rethink that input.
2. Another input Keith and Eliza suggested is the “cut plane” of the pimple. This approach seems equivalent to keeping the initial volume as a r.v. (where the initial volume will be some fraction of the pimple volume and the rest of the pimple volume that doesn’t flow will be absorbed into the elevation map), and adding a “down hill flow direction” (which depends on the angle of the cut plane) as a random variable. In this scenario we are assuming zero initial flux.
3. Apparently starting a large mass/volume as a heap is physically a bit unstable because of the tremendous potential energy such a heap would have. Q – If we are considering large volumes from cutting a pimple (assuming no flux), do we still have PE problems?
4. Another option here is to give the the flow from from the pimple (or heap) some initial velocity. With a large volume flow the initial speed might not be so important, but the initial direction of the velocity could be the angle we treat as a random variable. Apparently including an initial flux raises some numerical issues (see Keith for details) that we might be better off avoiding at this point.
5. Anything I’m leaving out? This seems like an important decision and we should look for some sort of consensus.

First problem

A reasonable goal for our first problem is to find the probability that the maximum flow height, h_{max} , at one specific interesting location, $x_{target} = (E_{target}, N_{target})$, exceeds some critical flow height, h_{crit} . First, let’s consider fixed friction coefficients and zero initial velocity (these will be issues to address later). Then we are looking for the *most likely* combination(s) of volumes and initial flow “angles” (this input needs to be clarified as discussed above under “more thoughts on model inputs”) that generate a flow such that $h_{max}(x_{target}) \geq h_{crit}(x_{target})$.

Naive idea

Ideally, we would sample volumes from Robert and Simon’s distribution, varying the initial coordinates uniformly (? or according to some known distribution for a given volcano) and use these values to initialize TITAN runs for a Monte-Carlo (MC) simulation and compute the probability, $P(h_{max}(x_{target}) \geq h_{crit}(x_{target}))$, directly.

Problems with this idea

1. It would take an unreasonable number of TITAN runs to find this probability with decent error bounds, even if the probability was fairly large.
2. Presumably, many TITAN runs in this scenario will give us no flow at the site of interest – obviously a waste of resources.
3. If $P(h_{max}(x_{target}) \geq h_{crit}(x_{target}))$ is small, we may only directly sample a few volumes large enough to get get a flow with h_{max} near h_{crit} and those flows may be “wasted” on initial angle that, say, send the flow down the other side of the mountain.

Slightly less naive idea: importance sampling an emulator

Emulating the surface that we wish to sample is attractive because (obviously) emulator evaluations are very cheap. Designing the emulator for this use probably won't be trivial. Working on the assumption that we are looking for the smallest volume flow to cause a h_{max} near h_{crit} at x_{target} , we can restrict the domain of the initial flow angles to be some interval centered around $\theta = 0$ where $\theta = 0$ "lines up" with the x_{target} – maybe to be determined with a course run. How large to make the range of angles will probably be problem specific (and informed by the geography, river valleys etc). We will also want the design space to cover a range of volumes that give us a range of $h_{max}(x_{target})$ around $h_{crit}(x_{target})$, say $h_{max} \in (\epsilon, 2h_{max}]$. Presumably this will take some tinkering to figure out an appropriate design domain – again, maybe we can find a reasonable volume range with a course run. Once we have a surface established, we can use importance sampling to

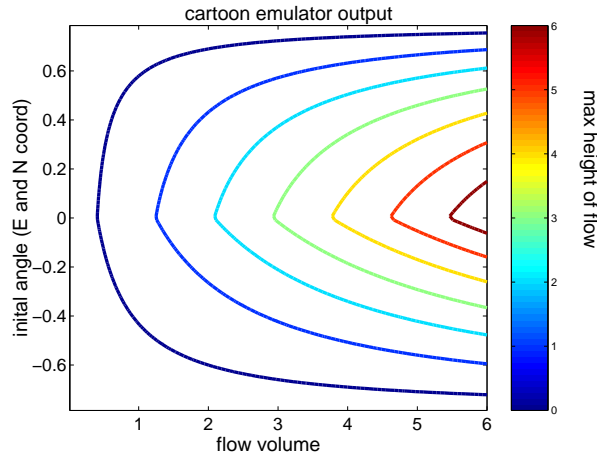


Figure 1: Cartoon of surface we wish to emulate and sample.

focus the samples in the region of the design domain.

How to proceed initially? (expanded based on last Thursday's meeting)

1. We could work backwards. That is, choose an x_{target} , a range of angles, and a range of volumes, then based on the h_{max} surface generated, choose an h_{crit} . (Maybe some runs already exist that might be suitable to use?) Clearly this is not the way to proceed for a realistic problem, but it seems like easier test of concept problem.
2. Begin with a coarse grid mesh to give us a sense of what the surface looks like. If target location wasn't specified a priori, pick one as described in (1).
3. Hopefully (2) will let us define an subset of the design space more likely to lead to the rare event we care about. Once we've chosen this subspace, we can put a design on it (Latin hypercube, Keith's Bayes linear adaptive design, etc.) and build an emulator of this interesting region. The purpose of building this emulator is to hone in on what might be a rather narrow region of the subspace where the rare events are *most likely to come from*. Because the emulator is easy to evaluate, we can do importance sampled Monte-Carlo (IS-MC), sampling from Robert and Simon's distribution and some subset of possible angles, and narrowly define the most important regions of design space without wasting a lot of TITAN runs.
4. If (3) is successful, we can do IS-MC or some sort of biased MCMC (this isn't quite clear to me) directly with runs of TITAN where inputs are specified by (3).
5. Other thoughts?