Multi-Model Comparison of Computed Debris Flow Runout for the 9 January 2018 Montecito, California Post-Wildfire Event

K. R. Barnhart1, R. P. Jones1, D. L. George2, B. W. McArdell3, F. K. Rengers1, D. M. Staley1, and J. W. Kean1

1 U.S. Geological Survey, Denver, CO, USA, 2U.S. Geological Survey, Vancouver, WA, USA, 3 Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland

Abstract  Hazard assessment for post-wildfire debris flows, which are common in the steep terrain of the western United States, has focused on the susceptibility of upstream basins to generate debris flows. However, reducing public exposure to this hazard also requires an assessment of hazards in downstream areas that might be inundated during debris flow runout. Debris flow runout models are widely available, but their application to hazard assessment for post-wildfire debris flows has not been extensively tested. Necessary inputs to these models include the total volume of the mobilized flow, flow properties (either inherent material properties or calibration coefficients), and site topography. Estimates of volume are possible in post-event (“back calculation”) studies, yet before an event, volume is an uncertain quantity. We simulated debris flow runout for the well-constrained 9 January 2018 Montecito event using three models (RAMMS, FLO2D, and D-Claw) to determine the relative importance of volume and flow properties. We broke the impacted area into three domains, and for each model-domain combination, we performed a numerical sampling study in which volume and flow properties varied within a wide, but plausible range. We assessed model performance based on inundation patterns and peak flow depths. We found all models could simulate the event with comparable results. Simulation performance was most sensitive to flow volume and less sensitive to flow properties. Our results emphasize the importance of reducing uncertainty in pre-event estimates of flow volume for hazard assessment.

Plain Language Summary Changes in soil properties after fire mean rain can more easily erode the surface and initiate debris flows—mixtures of water, soil, and rock that rapidly move from steep source areas to downstream regions. The combination of frequent fire, steep landscapes, dense population, and intense rain makes southern California prone to debris flows. The scientific community has established methods linking debris flow likelihood and volume to rainfall intensity and burned area characteristics. We presently lack a basis for understanding how flow volume, flow material properties, representation of flow physics, and site topography all combine to produce an inundation hazard. We begin to address this by using observations from the 9 January 2018 Montecito, California, debris flow event to test three candidate models for debris flow inundation. We found that all models can simulate the event and that the largest source of uncertainty in inundated area is the flow volume.

1. Introduction

Wildfire alters the properties of surface soils, leading to increased surface water runoff, sediment yield, and probability of debris flows during intense rainstorms (e.g., Cannon et al., 2001; Ice et al., 2004; Moody et al., 2013; Parise & Cannon, 2012; Santi & Rengers, 2020). Debris flows are water-laden slurries of soil and rock that move rapidly through channels in steep landscapes. When debris flows exit their steep upland source basins and run out onto urbanized debris flow fans or floodplains, they present a hazard to human life, infrastructure, and property. Southern California is especially exposed to this hazard because of the combination of dense population adjacent to steep topography and the capacity for high intensity rainfall. The current and future state of precipitation intensity-duration characteristics and fire frequency result in high probability of debris flows in southern California every year and major events (damaging more than 40 homes) every 10–13 years (Kean & Staley, 2021).

Given the size, burn severity, soil characteristics, and topography of a burned basin, there is an established basis for estimating whether a design storm will produce a debris flow (Staley et al., 2016, 2017) and how large such
Understanding inundation hazard is complex for a number of reasons. First, the physics of debris flows is different from that of water flows. Debris flows may have peak discharges that are many times greater than floods. Debris flows may clog channels and cause avulsion or deposit levees which steer flow. These characteristics make simulating runout dynamics difficult. Second, it is not clear which pre-event unknowns (e.g., the spatial extent and intensity of a storm, the plausible range of debris volume produced, the material properties of a flow) have the largest impact on the inundation patterns. In contrast to forensic back-calculations after an event, hazard assessments are conducted before a hypothetical event. Some of these unknowns may be constrained by field observations (e.g., material properties), and others are unconstrained before a meteorological forecast is available (e.g., spatial extent and intensity of a storm). Further, the relative contribution of event characteristics (e.g., event volume, material properties) and intrinsic site properties (e.g., alluvial fan morphology) to the overall hazard is not constrained. Therefore, understanding the relative contribution of each of these uncertain inputs to predicted quantities of interest, such as which areas may be inundated, is useful for hazard assessments.

Many debris flow inundation models aim to simulate inundation patterns and flow characteristics given a volume of material, material properties, and information about site topography (see review by McDougall, 2017). These models are typically either empirical—implementing observed relationships between debris flow volume and inundation characteristics—or numerical—implementing conservation equations for material flow under the influence of gravity. Inundation models are commonly tested against well-documented field data in which information about flow volume and physical properties is known (e.g., Bertolo & Wieczorek, 2005; Di Santolo & Evangelista, 2009; Fan et al., 2017; Frank et al., 2017; Hussin et al., 2012; Rickenmann et al., 2006; Schraml et al., 2015). In these studies, it is most common to compare one or more simulation codes and to assess model sensitivity to flow property inputs under the condition that volume is known. A more limited subset of these back-calculation studies explores the influence of flow volume on model performance (e.g., Frank et al., 2017; Melo et al., 2018; Schraml et al., 2015). In a pre-event context, however, there is limited information regarding both flow volume and flow properties. Use of runout models in post-wildfire applications is even more limited (Bessette-Kirton et al., 2019; Cannon et al., 2009; McCoy et al., 2016; Melo et al., 2018).

We examined three candidate inundation models in the context of the 9 January 2018 Montecito event (Kean et al., 2019b; Lancaster et al., 2021; Oakley et al., 2018). The study focused on a single rainfall event, and we considered all five of the most destructive runout paths. Based on shared characteristics, these five runout paths were grouped into three domains. For each combination of model and domain, we evaluated many parameter sets to assess the relative role of volume and flow properties in simulation performance. Comparing the best simulations from each model served as a test of which model representation of debris flow physics and initialization may better represent runout and inundation. Comparing the three domains represented an initial step in assessing the role of source basin and fan morphology— intrinsic site characteristics— on inundation hazard and on the transferability of flow property values between domains.

Each of the three models represents the initialization of the flow volume in a different way. Two use inflow hydrographs in which a time-discharge curve is a required input, and the third permits the arbitrary distribution of material throughout the domain. We refer to these different requirements as differences in model epistemology or what the model expects the user to know in advance of using it. Because our study design compared models with different epistemologies, we explored how these differences manifest in the questions a user is able to ask with each type of model.

Section 2 describes the 9 January 2018 event and presents estimates of the total volume of moving sediment and water derived from each of the three source basins. We establish the scope (Section 3) and design of our numerical experiment (Section 4). Because we compared three models that vary in their governing equations and initial conditions and we interpreted best-fit parameters from these governing equations, we provide a comprehensive overview of these governing equations in Text S1 in Supporting Information S1. Each model requires a different approach to initialization—although every attempt was made to reduce these differences. We describe model initialization and parameter input values in Section 5. Formal assessment of simulation performance required an
expression of goodness of fit, in this case based on matching field observations of inundation extent and peak flow depth (Section 6). Section 7 describes the results and discusses the implications for pre-event uncertainty in volume, the role of site morphology, and intermodel comparison.

2. 9 January 2018 Debris Flow Event

Between 4 December 2017 and 12 January 2018, the Thomas Fire burned 114,078 ha in Santa Barbara and Ventura Counties, CA (CAL FIRE, 2018). Based on a design rainstorm with 15 min rainfall intensity \( I_{15} \) \([\text{L T}^{-1}]\) (Dimensions introduced are given as M, mass; L, length; T, time; - , nondimensional) of 24 mm hr\(^{-1}\) the Santa Ynez Mountains to the north of the municipality of Montecito, CA, were identified as having high debris flow hazard (U.S. Geological Survey, 2018). On 9 January 2018, a high-intensity rainfall event resulted in rainfall rates ranging from \( I_{15} = 84 \text{ mm hr}^{-1} \) to 99 mm hr\(^{-1}\) (Oakley et al., 2018). This storm produced debris flows throughout the western portion of the area burned by the Thomas Fire (Kean et al., 2019b; Lancaster et al., 2021; Lukashov et al., 2019), with the greatest impacts in Montecito (Kean et al., 2019b, and Figure 1). Debris flows originated in the Santa Ynez Mountains and ran out onto the urbanized area of coalesced alluvial fans. The debris flows roughly followed the major drainages of Montecito Creek, Oak Creek, San Ysidro Creek, Buena Vista Creek, and Romero Creek. Debris overflowed the active channels at and upstream of structures such as culverts and bridges (Kean et al., 2019b). The storm mobilized a debris flow volume of \( \sim 680,000 \) m\(^3\) (Kean et al., 2019b, and Figure 1). This event killed 23 people, damaged 558 structures, closed U.S. Highway 101 for 13 days, and caused infrastructure damage in Montecito and neighboring Carpinteria, CA.

Kean et al. (2019b) constrained the event timing based on rain gage data, security camera footage, emergency telephone calls, and seismic signals. Rainfall peaked just before 3:45 a.m. (local time); a security camera recorded the initial debris flow front exit Cold Spring Creek (a tributary to Montecito Creek) at 3:49 a.m.; and the seismic signal at station CI.QAD.EHZ (located in the southeast portion of Montecito) peaked at 4:07 a.m. Based on these observations, we concluded that the majority of the flow transited the coalesced alluvial fan in about 20 min. We converted this duration into a rough estimate of debris flow velocity using the distance from mountain front to ocean along Montecito Creek (4.3 km). The calculated velocity of 13 kph (3.6 m/s) is within the error of super-elevation-based velocities observed in the source basins (4.5 ± 50% m/s, Kean et al., 2019b).

In this study, we split the coalesced alluvial fan into three regions, each of which encompasses both an upstream debris flow source basin and a downstream runout domain. The runout domains were constructed by grouping major adjacent flow paths and are depicted in Figure 1b. The source basins are depicted in Figure 1c. Each runout domain has multiple upstream source basins, fed by one or more major creeks. The Montecito Domain consists of the east and west forks of Cold Spring Creek and Hot Springs Creek (which join to form Montecito Creek), the San Ysidro Domain contains both San Ysidro and Oak Creeks, and the Romero Domain includes Romero and Buena Vista Creeks.

The topographic setting of each domain is slightly different owing to elevated topography created by the Mission Ridge fault zone (Kean et al., 2019b). The drainage area of source basins for the Montecito, San Ysidro, and Romero domains are 12.1, 8.7, and 7.7 km\(^2\), respectively. The simulated runout domains shown in Figure 1b are 7.0, 5.7, and 5.6 km\(^2\), respectively.

The source basin upstream of the Montecito, San Ysidro, and Romero runout domains yielded estimates of 231,000, 307,000, and 141,000 m\(^3\) of sediment \( V_s \) \([\text{L}^3]\), respectively (Kean et al., 2019b). Because depth observations were evenly distributed, the volumes were calculated by these authors using a simple mean of sediment deposit depths multiplied by the inundated areas. These estimates are inherently uncertain, and we did not attempt a formal estimate of either of these quantities. However, to reflect the uncertainty on these and the volume estimates for water presented below, we represent factor of two (−50%/+100%) uncertainty bounds.

Field evidence suggested that the solid volume fraction was higher in the San Ysidro and Romero domains than in the Montecito domain (Kean et al., 2019b). No direct observations of solid volume fraction exist, but comparing the mean sediment deposit depth \( \bar{h}_s \) \([\text{L}]\) with mean flow marker depth \( \bar{h} \) \([\text{L}]\) indicated that the solid volume fraction of the San Ysidro domain was likely higher than the other two drainages. Values for \( \bar{h}/\bar{h}_s \) range from 0.14 for Montecito Creek, 0.20 for Oak Creek, 0.38 for San Ysidro Creek, 0.11 for Buena Vista Creek, and 0.23 for Romero Creek (Kean et al., 2019b).
For each of the three domains we used field observations of rainfall and the total volume of sediment deposited to estimate the total flow volume. To estimate the water volumes, we used rainfall derived from the Doulton Tunnel and KTYD gages (locations in Figure 1) and partitioned rainfall into infiltration and runoff using the method of Chu (1978, our Figure 2). Modeling-derived basin-averaged infiltration rates for similar sites are 1–10 mm hr\(^{-1}\) (Rengers et al., 2016), and post-event field measurements were 20 mm hr\(^{-1}\) (Kean et al., 2019b). We used an infiltration rate of 10 mm hr\(^{-1}\), reflecting the highest modeling-derived value. We expect that the field measurements, taken after the event, overestimated the event infiltration rate.
The partitioning analysis yielded runoff depths of 28.4 and 21.0 mm for Doulton and KYTD, respectively. We used the average value of 24.7 mm to calculate a total water volume for each domain: 300,000, 215,000, and 191,000 m$^3$ for Montecito, San Ysidro, and Romero domains, respectively. The rain gage records indicate a rainfall gradient, with higher rainfall in at the eastern rain gage, Doulton. We do not include this gradient in our water volume calculations, but estimate it with the quotient 3.7 mm/24.7 mm, or 15%. As this is well within our factor of two uncertainty bounds for event volume, we do not consider this source of uncertainty further.

We combined the water and sediment volumes to obtain an estimate of the flow volume and solid volume fraction for each region: 531,000 m$^3$ and 0.435 for Montecito; 522,000 m$^3$ and 0.588 for San Ysidro; and 332,000 m$^3$ and 0.425 for Romero. Bessette-Kirton et al. (2019) estimated a total volume of 495,000 m$^3$ for San Ysidro alone based on an assumed solid volume fraction of 0.60, in agreement with our estimates.

We compared the sediment volume estimates for each region with a range of sediment volumes predicted by the empirical relationship of Gartner et al. (2014), which is currently used for post-wildfire debris flow hazard assessments (Figure 3). The field-based sediment volume estimate for each basin lies at or below the expected value from the Gartner et al. (2014) relationship. Additionally, this figure illustrates the magnitude of pre-event uncertainty in volume based on currently available methods to estimate it—conditional on an accurate rainfall intensity forecast. The prediction uncertainty interval for one standard error nearly spans an entire order of magnitude.

In addition, we placed the magnitude of this event in the context of prior work establishing the relationship between observed debris flow volumes $V$ [L$^3$] and inundated planimetric area $B$ [L$^2$] from a general compilation (Griswold & Iverson, 2008) and a post-wildfire compilation of the 2003 Christmas Day storms impacting the Grand Prix and Old Fires in southern California (Bernard et al., 2021) (Figure 4). Griswold and Iverson (2008) estimate a relationship of $B = 20 V^{2/3}$ and argue for setting the exponent to 2/3 based on the

---

**Figure 2.** Observed rainfall intensity (a) partitioned into cumulative runoff and (b) infiltration at the Doulton Tunnel and KYTD station sites using an infiltration rate of 10 mm hr$^{-1}$.

**Figure 3.** Relationship between 15-min rainfall intensity $I_{15}$ and mobilized sediment volume $V_s$ predicted by the empirical relationship of Gartner et al. (2014). Black line reflects the expected value, and the gray region reflects ±1 standard error. White diamonds show observed $I_{15}$ and $V_s$ for each domain. Vertical error bars reflect a factor of two uncertainty range for $V_s$. 

---
Figure 4. Relationship between flow volume $V$ and inundated planimetric area $B$. Comparison between data and statistical model fits from Griswold and Iverson (2008) and Bernard et al. (2021).

physical relationship between area and volume. An exponent of 2/3 yields dimensionless (and thus comparable) coefficients, and Griswold and Iverson (2008) found equivalent statistical fits using either a free or fixed exponent. Bernard et al. (2021) estimated a relationship of $B = (7.4 \text{ meters}^{1.22}) V^{0.81}$ and found this fit was substantially better when the exponent was permitted to vary; a fit to their data with an enforced exponent yielded $B = 28 V^{2/3}$. This larger prefactor (28 vs. 20) indicates that post-wildfire debris flows inundate a greater area for their volume than other debris flows. Calculating the prefactor ($B/V^{2/3}$) for each of the three domains yields a range of 150–170 for the central volume estimate and 100–270 for the factor of two uncertainty range.

For the volume mobilized, the 9 January 2018 event inundated a larger than expected area based on either empirical fit. The event sits outside of the 95% prediction confidence level under the Griswold and Iverson (2008) fit and the 80% level under the Bernard et al. (2021) fit. This observation is consistent with the simulated lack of inundation of this event by LaharZ (Bessette-Kirton et al., 2019), which uses the Griswold and Iverson (2008) scaling relationship.

3. Model Selection Scope

We considered three models for simulating debris flow inundation that vary in their representation of debris flow physics as well as in their approach to initialization. Our intention is not to be comprehensive in our model selection, as there are many other suitable models. Instead, our goal is to compare models that span a range of flow representation. By comparing simulation performance across these models, we present a formal test of the role of model physics in simulation results.

Two of these models are single phase, meaning they represent the complex interactions between fluid and solid in debris flows (e.g., Iverson, 1997) as a single phase of flowing material with a specified relationship between
shear stress and strain rate. We consider RAMMS (Christen et al., 2010) and FLO-2D (O’Brien et al., 1993) because they use different shear stress versus strain rate relationships and because they are each widely used. The final considered model is D-Claw (George & Iverson, 2014; Iverson & George, 2014), which represents both fluid and solid fractions as a quasi-two-phase flow.

There are many characteristics of debris flows that none of the considered models represent. For example, three dimensional (3D) effects such as culverts are not represented in depth-averaged models. However, by comparing these models we were able to assess which model’s representation of debris flow physics were necessary for simulating this event.

We do not consider other single-phase simulation codes (e.g., DAN3D, Hungr & McDougall, 2009), other multi-phase simulation codes (r.avaflow, Mergili et al., 2017; Pudasaini, 2012; Pudasaini & Mergili, 2019), or empirical scaling relationship-based flow runout frameworks (e.g., LaharZ, Schilling, 1998; Iverson et al., 1998) in order to keep the computational scope of this work manageable (see McDougall, 2017, for a review of runout modeling). In addition, the three considered models support our aim of model intercomparison and permit comparison of single versus quasi-two-phase flow, and comparison of two shear-stress versus strain rate relationships.

4. Numerical Experiment Design

We conducted a sampling study in which we defined the plausible parameter space for each model and sampled it to explore the relationship between model specification and model performance metrics. We used a Latin hypercube sampling (LHS) design implemented in the Dakota software (Adams et al., 2020) because of its efficient space-filling qualities (McKay et al., 1979; Swiler & Wyss, 2004). We evaluated 100Np samples for each model and each domain, where Np is the number of input parameters used to specify each model (Np = 3, 5, 4 for RAMMS, FLO-2D, and D-Claw, respectively). Each sample resulted in an independent numerical simulation. This design used a sufficient number of samples for screening the parameters that are most important for each model and identifying the regions of parameter space that have good model performance (Pianosi et al., 2016).

The multi-model, multi-domain, LHS study supported model intercomparison, domain intercomparison, and sensitivity analysis in which we interrogated the relationship between model input parameters and model performance. We considered three components of model performance: (a) matching the observed inundation patterns, (b) overestimate of peak flow depth, and (c) underestimate of peak flow depth (Section 6). We grouped the model inputs into two categories: those that control the total flow volume and those that control the flow properties. We addressed the following questions:

1. What are the relationships between the three components of model performance?
2. What are the relative roles of flow volume and flow properties on each element of model performance?
3. After controlling for volume, are there parts of the parameter space that provide systematically better model performance?
4. How good are the best simulations?

We used this initial set of simulations to address the above four questions in Sections 7.3–7.6. Natural extensions were to ask “how transferable are the input values used in the best simulations?” and “how comparable are the best simulations to default, expert opinion-based input values?” This first question addressed how unique the best-fit inputs are between the three considered domains. The second question addressed how the best-fit model results compare with the type of results one might expect from a pre-event hazard assessment (https://landslides.usgs.gov/hazards/postfire_debrisflow/).

To address these final questions, we supplemented our LHS study with an additional set of simulations (Section 7.7). For these simulations, we specified the total flow volume and solid volume fraction based on the field observations for each domain. We identified the best input parameter set for each domain, conditional on using a volume and solid volume fraction that is close to the field observation. We then swapped the best-fit physical property inputs between the domains. Finally, we ran one simulation for each combination of domain and model using expert opinion-based input parameters.
5. Model Specification

All three models required a specification of initial topography, the total flow volume, and flow property input parameters. Each model supports a different set of physical processes and representation of the environment. Our use of each model was designed to make the simulations as comparable as possible in order to compare the role of governing equations; thus, we did not use all the features of each model. Example model features we did not use include the FLO-2D components that represent storm drains and support a spatially variable Manning’s $n$ value, and RAMMS capabilities to simulate entrainment of material or spatially variable flow property coefficients.

5.1. Input Parameters

Each of the three models solves a system of depth-averaged conservation equations, supplemented by constitutive relations. Governing equations for each of the three models are presented in Text S1 in Supporting Information S1. Here, we summarize the key inputs for each model.

RAMMS calculates the frictional deceleration following the Voellmy friction relation (Salm, 1993; Salm et al., 1990). The flow properties in RAMMS are controlled by two coefficients $\mu_r [-]$ and $\xi [L T^{-2}]$ (Christen et al., 2010). FLO-2D uses a “quadratic” relationship introduced by O’Brien and Julien (1985). This shear-stress versus strain rate relationship is specified with the following three coefficients: $Y_S [M L^{-1} T^{-2}], \mu_r [M L^{-1} T^{-1}]$, and $K [-]$. Additionally, FLO-2D uses a Manning’s $n$ value $[T L^{-1/3}]$. Thus, including specification of volume, RAMMS uses three inputs, and FLO-2D uses five.

Because D-Claw represents both solid and fluid volume fractions, it requires specification of the initial solid volume fraction $m_0 [-]$. In D-Claw, the solid volume fraction $m [-]$ evolves through space and time along with the flow depth and velocity. A hydraulic permeability $k [L^2]$ evolves as a function of $m$ and uses an initial value $k_0$. Following prior applications of D-Claw, we varied only $m_0$, $k_0$, and the basal friction angle $\phi_{bas}$ [-] (George & Iverson, 2014; Iverson & George, 2014, 2016). This decision was further supported by the dimensional analysis of the governing equations presented in Iverson and George (2014), which reduced the parameter space to six nondimensional inputs. We transformed the inputs into these nondimensional parameters and present them as $\phi_{net}$, the difference between $m_0$ and the critical state volume fraction $m_{crit}$, and the timescale ratio for downslope debris motion and pore-pressure diffusion $k’$, originally defined in Iverson and George (2014), their Equation 5.10:

$$k’ = k_0 (L/g)^{0.5} / \alpha \mu_r H^2,$$

where $\alpha$ is the debris’ elastic compressibility and $\mu_r$ is the effective shear viscosity of the pore fluid. Calculation of $k’$ required specification of a characteristic length scale $L$ and thickness scale $H$, and we use 6,000 m and 3 m, respectively, to reflect the scale of this event. D-Claw used a total of four inputs.

5.2. Discretization, Topography, and Simulated Duration

For all three models, we used a spatial discretization of 5 m. D-Claw permits adaptive mesh refinement and was specified to use a 5 m cell size anywhere debris flow material was present. Where debris flow material was not present, cells could remain coarse, up to 1,000 m in width. The narrowest and most entrenched portions of the mainstem channels traversing the coalesced alluvial fan are 15–20 m wide, whereas the inundated areas commonly reached over 100 m wide. Our discretization provided multiple grid cells within the narrowest portions of the confined channel (Wehrli, 2019). All model simulations used the same topography. Pre-event light detection and ranging (lidar) acquired by Harris Geospatial in 2015 and purchased by the U.S. Geological Survey (USGS), covers most of the considered source and runout areas. These data are licensed and thus not publicly available. In the portions of upper Romero Creek and the eastern area of the coalesced alluvial fan where the pre-event lidar does not have coverage, we used post-event lidar collected on 17 January 2018 and publicly available through the USGS 3D Elevation Program (U.S. Geological Survey, 2017). For the purposes of this study, both lidar data sets were processed from raw point cloud to a 5 m bare-earth digital elevation model using LASTOOLS (rapidlasso GmbH, 2020). The topographic surface used in modeling thus represents the topography with buildings and structures removed.
The elapsed real time of running each model for a desired simulated duration depended on the discretization, and using a 5 m grid cell spacing made the numerical experiment computationally feasible. FLO-2D and RAMMS ran on a Windows 10 computer (2 × 8 core, Intel Core i9-988H 2.30 GHz). The average simulation time varied depending on the model and domain: 2.46 SU (1 SU = 1 core-hour), 1.45 SU, and 1.37 SU for RAMMS; 1.94, 1.68, and 2.04 SU for FLO-2D for the Montecito, San Ysidro, and Romero domains, respectively. D-Claw ran on the USGS High Performance Computer Denali and required runtimes of 894, 630, and 907 SU, respectively (U.S. Geological Survey Advanced Research Computing, 2021). Denali uses 2 × 20 core 2.4 GHz Intel Xeon Gold 6148 processors (Skylake).

All models use adaptive time stepping to keep simulations numerically stable. RAMMS used a Courant-Freidrich-Levy (CFL) criterion of 0.5. FLO-2D permits the user to specify the CFL criterion, which was set to 0.6 for this application. D-Claw requires specification of desired and maximum CFL criteria, set to 0.25 and 0.5, respectively.

The duration of simulated time was guided by the field and seismic observations, which constrain event timing to a total of two hours. Both RAMMS and D-Claw are able to stop before the end of the specified simulation time, if the moving momentum has fallen below a specified threshold of maximum moving momentum. This option was used with the threshold set to 5%.

5.3. Flow Volume Representation and Initialization

The Montecito, San Ysidro, and Romero domains are each fed by multiple source basins. For example, the Montecito domain is fed by the east and west forks of Cold Spring Creek, Hot Springs Creek, and a number of smaller basins. RAMMS and FLO-2D use inflow hydrographs at the mouth of each source basin. We do not use the RAMMS “block release” capabilities. At these locations, we specify a unique discharge-time relationship for each simulation. Following the default behavior of RAMMS, for both FLO-2D and RAMMS, we use a triangular hydrograph with the peak discharge (which implies hydrograph duration) calculated based on the total volume of water and sediment $V_{e}$ [L$^{3}$] after Rickenmann (1999, his Equation 1). The timing of peak discharge was set to halfway through the hydrograph duration. Both RAMMS and FLO-2D use inflow hydrographs, but the way in which the inflow cells are specified differs slightly such that it was not possible to use the same set of inflow model grid cells across RAMMS and FLO-2D. We used the closest set of inflow cells in each model to strive for consistency.

In contrast, D-CLAW considers the potential mobilization of static material distributed in an arbitrary pattern over the entire domain. At the onset of a simulation, this material may become unstable and flow. The distinction between the inflow hydrograph approach and the distributed material approach represents a difference in model epistemology. Using an inflow hydrograph requires that a user knows the volume of material that passes the inflow location, but also means that a user can ensure that a specified volume of material will move onto the fan. In contrast, the distributed material approach does not require the user to know the shape of the hydrograph—but it also does not guarantee that a given flow volume will pass the fan apex. Accordingly, each initialization approach may be able to address questions that the other cannot. One such example is our analysis of how D-Claw input parameters impact the amount of volume that passes beyond the inflow hydrograph points (Section 7.1).

For the purposes of this application, the total flow volume was distributed in D-Claw on and adjacent to the main channels in each source basin, with a constant thickness (Figure 5). A buffer of 25 m to either side of the main channels was used to delineate the initialization region. This initialization was not intended to represent the initiation processes of the actual event, in which material was sourced from the hillslopes and channels. Initial tests indicated that the resulting hydrograph at the fan apex was not overly sensitive to the details of initialization and that the shapes of hydrographs constructed using the Rickenmann (1999) scaling relationship are comparable with those simulated by D-Claw (Figures S1–S3 in Supporting Information S1). Because this initialization scheme does not guarantee the entire flow volume runs out past the fan apex, we calculated the “effective” volume $V_{f}$, or the volume that passed into the runout domains (Figure 1b). We consider $V_{f}$ for D-Claw when comparing between models because it is comparable to $V_{e}$ as used in the other two models (for RAMMS and FLO-2D, $V_{f} = V_{e}$).

For all three models, we needed a method to partition the total volume in each simulation into the volume originating from each source basin (or volume exuded at each inflow hydrograph point). We partitioned the total volume for a given simulation based on the USGS hazard assessment for the Thomas Fire (U.S. Geological Survey, 2018), which implements the empirical debris flow volume model of Gartner et al. (2014). This empirical model relates total volume of produced debris with the 15-min rainfall intensity, the relief of the basin, and the...
area burned at moderate to high severity. Other approaches for predicting volume exist; however, this empirical relationship is the most widely used and most successful at predicting post-wildfire debris flow volume in this region (Rengers et al., 2021).

The total flow volume $V_t$ is the only input shared across all three models. We used a large volume range in this study in order to understand how inundation characteristics vary as a function of flow volume. Our range spanned 50,000–5,000,000 m$^3$ and was chosen to reflect plausible total volumes under design storms which have $I_{15}$ rates between 12–110 mm hr$^{-1}$ (Figure 3). Here, the lower limit reflects the smallest design storm produced by USGS post-wildfire hazard assessments, and the upper limit is a slightly larger rainfall intensity than was observed in the event. We expect that the largest and smallest volumes will not represent the event dynamics well; however,
exploring the properties of each model across volumes provides insight on the relative role of flow properties and volume in producing inundated area.

### 5.4. Input Parameter Ranges

Each of the three models required a set of input parameters that influence the flow properties (Table 1). Some flow properties are best considered as calibration parameters, whereas others are intended as inherent material properties. We will refer to all as input parameters. Because we present a sampling study within a predefined range, our results are inherently conditional on the specified ranges. However, in specifying each input parameter range we sought the widest, yet still physically plausible range. Such a range was chosen to understand the relative contribution of plausible physical property variations in comparison with our large variation in volume. Only the range of \( m_0 \), the initial solid volume fraction in D-Claw was directly based on field observation and lab measurements.

**Table 1**

| Variable Symbol | Units | Min | Max | Expert Opinion |
|-----------------|-------|-----|-----|----------------|
| **A. RAMMS**    |       |     |     |                |
| \( \log V_t \) | m\(^3\) | 4.7 | 6.7 | M: 5.72, S:5.71, R: 5.52 |
| \( \mu_r \)    | -     | 0.001 | 0.1 | 0.05 |
| \( \xi \)      | m.s\(^{-2}\) | 10 | 2,000 | 1,000 |
| **B. FLO-2D**   |       |     |     |                |
| \( \log V_t \) | m\(^3\) | 4.7 | 6.7 | M: 5.72, S:5.71, R: 5.52 |
| \( \log \mu_f \) | Pa.s | –2 | 1.7 | 1 |
| \( \log Y_s \) | Pa    | –1 | 2.6 | 2 |
| \( \log K \)   | -     | 1.38 | 3.7 | 3.36 |
| \( n \)        | s m\(^{-1/3}\) | 0.006 | 0.27 | 0.027 |
| **C. D-Claw**   |       |     |     |                |
| \( \log V_t \) | m\(^3\) | 4.7 | 6.7 | M: 5.72, S:5.71, R: 5.52 |
| \( m_0 \)      | -     | 0.42 | 0.62 | M: 0.43, S:0.58, R: 0.42 |
| \( \log k_0 \) | m\(^2\) | –12 | –8 | –11 |
| \( \phi_{bed} \) | - | 30 | 45 | 36 |

Note. M = Montecito, S = San Ysidro, R = Romero domain.
Our application of D-Claw permitted \( k_0, m_s, \) and \( \phi_{\text{crit}} \) to vary. We set \( m_{\text{sed}} \) in all simulations to 0.64. The observed porosity of 37.7% (Kean et al., 2019b) provided an upper bound for the initial solid volume fraction \( m_s = 0.62 \). Because the mobile material was sourced from hillslope erosion, the effective initial solid volume fraction may have been smaller. We considered a range that extends to 0.42, reflecting the water and sediment volume estimates made in Section 2. The basal friction angle varied from \( \phi_{\text{bed}} \) in the range of 30–45°. Finally, the initial permeability \( k_0 \) varied from \( 10^{-14} \) to \( 10^{-8} \), reflecting previous estimates of plausible values (Iverson & George, 2014, 2016) and field observations (Kean et al., 2019b). The basal pore pressure was initialized to hydrostatic pressure.

### 6. Simulation Performance Assessment

We used our LHS study to determine the best-fit input parameters for each model that minimized a misfit metric. In this section, we describe how model results were reduced into this misfit metric. We considered three model outputs as metrics of model-data comparison and combined them into one cumulative misfit metric. The first of these is a measure of the misfit between observed (true) and simulated inundation extent. Model grid cells can be classified into true positive (TP), false positive (FP), false negative (FN), and true negative (TN). These four quantities are collectively called the “confusion matrix” and represent a summary of binary classification performance. We used a threshold of \( h > 0.5 \) m to classify inundated areas and justify this value in Section 7.2. Heiser et al. (2017) introduced a metric \( \Omega_f \), appropriate for use in assessing debris flow inundation

\[
\Omega_f = \frac{TP - FP - FN}{TP + FP + FN},
\]

which ranges from −1 (worst performance) to 1 (best performance). This metric has similar qualities to the threat score in that it neglects the TN area. Because \( TP + FP + FN + TN = 1 \), three of the four elements contain all information from the confusion matrix. TN is the most natural element to omit because it is not directly associated with either observed or simulated inundation, instead it reflects the size of the computational domain or area of interest. For convenience and comparability, we modify \( \Omega_f \) so that it scales between 0 (best performance) and 1 (worst performance)

\[
\Omega_{f,\text{tn}} = \frac{1 - \Omega_f}{2}.
\]

\( \Omega_{f,\text{tn}} \) is equivalent to one minus the threat score. In this analysis, we use the inundation extent mapped by Kean et al. (2019a) and clipped to the runout domains depicted in Figure 1b and downstream of the inflow locations used for FLO-2D and RAMMS.

The final two elements of model-data comparison are based on measured peak flow depths. For simplicity, we did not consider an additional metric based on the fragility curve-based velocities reported by Kean et al. (2019b). We used peak flow depth instead of observed deposit depth because none of the considered models directly simulated deposition; thus, peak flow depth was the most comparable simulated equivalent available. Furthermore, peak flow depth is more relevant for evaluating risk to infrastructure than deposit depth.

For each individual observation of peak flow depth \( d_{oi} \) measured in the field (Kean et al., 2019a, and Figure 1), the model grid cell containing that measurement was identified, and the corresponding modeled peak flow depth was extracted \( d_{m,i} \). The flow depth residual \( r_i \) was calculated as

\[
r_i = d_{oi} - d_{m,i}.
\]

We did not use the root mean squared error between the measured and modeled maximum depth because it combines the misfit from under- and overestimation. Because our analysis varies volume over a wide range, we expect underestimation to dominate at low volumes and overestimation to dominate at high volumes. However, it would be useful to know whether these misfit metrics trade off exactly, or if there are material property parameters that allow for a change in one without the other. Accordingly, we keep under- and overestimation separate. Additionally, the TP, FP, and FN areas are correlated with over- and underestimation. Peak depth observations are only located where the debris flow inundated (TP and FN). Overestimates are only found in the TP areas, where simulated and observed debris flows overlap. In contrast, underestimates can come from both the TP and FN areas, with all debris flow observations from the FN area contributing to the underestimate.
We defined two new metrics, the normalized sum of overestimates $\Delta_o$ and normalized sum of underestimates $\Delta_u$, defined as

$$
\Delta_o = \frac{1}{\sum_{m}^N \sum_i (||r_i||)} \sum_i^N \begin{cases} 
||r_i|| & \text{if } r_i > 0 \\
0 & \text{if } r_i \leq 0 
\end{cases}
$$

(5)

$$
\Delta_u = \frac{1}{\sum_{m}^N \sum_i (||r_i||)} \sum_i^N \begin{cases} 
||r_i|| & \text{if } r_i < 0 \\
0 & \text{if } r_i \geq 0 
\end{cases}
$$

(6)

where $N_m$ is the number of depth measurements and the prefactor normalizes the result to the sum of all residuals. $N_m = 223, 143, 258$, for the Montecito, San Ysidro, and Romero domains, respectively. Perfectly matching observations yields a value of 0 for both $\Delta_o$ and $\Delta_u$. The worst possible performance is bounded at 1 for $\Delta_u$, whereas it is unbounded for $\Delta_u$. For the sake of comparability between $\Omega_Tm$, $\Delta_o$, and $\Delta_u$, we define a capped version of $\Delta_o$

$$
\Delta_{o,c} = \begin{cases} 
\Delta_o & \text{if } \Delta_o < 1 \\
1 & \text{if } \Delta_o \geq 1 
\end{cases}
$$

(7)

Thus, poor performance indicated by high $\Delta_o$ reflects simulated peak flow depth being higher than observed, and high $\Delta_u$ reflects simulated peak flow depths being smaller than observed. In our discussion of the results we consider how $\Omega_Tm$, $\Delta_o,c$, and $\Delta_u$ are related (Section 7.3). We also calculated a final metric that combines these three elements. We weight the inundation extent and maximum depth elements equally and thus constructed our cumulative metric, $C_m$, as

$$
C_m = \frac{1}{4}(2\Omega_Tm + \Delta_{o,c} + \Delta_u).
$$

(8)

The cumulative metric $C_m$ varies between 0 and 1; it is used to identify the best performing simulations and to assess how simulation performance degrades as parameter values are changed between domains.

7. Simulation Results

The LHS permits a nearly infinite set of options for examining the simulation results. We constrained our analysis and looked first at what controls the volume that leaves upstream basins in D-Claw simulations and the relationship between volume and inundated area. Next, we examined the relationship between the considered outputs, the influence of volume on these outputs, and the influence of flow property inputs on the combined metric $C_m$. Finally, we explored the best simulations and identify the extent to which performance degrades when the best-fit inputs from one domain are used on another.

7.1. Volume Leaving Source Areas in D-Claw

The initialization of D-Claw with mobile material spread along and immediately adjacent to the source basin channels permits examination of how initial volume and material properties influence how much material exited the source basin (Figure 6). Given the D-Claw governing equations and prior work by Iverson and George (2016), we would expect that simulations with high $k'$, $\phi_{in}$, and $m_0 - m_{crit}$ would stop more quickly and would have smaller exit percentages ($100 \frac{|V|}{V_i}$). We note that our numerical experiment only considered inputs with $m_0 - m_{crit} < -0.02$; thus, all simulations had initially contractive, mobile flows.
Across all domains, the higher the total volume, the higher the percentage of material that exited the source basin. The distribution of exit percentages was bimodal, especially at high volumes, in that either most of the material exited the source basin, or little exited. To analyze how flow properties influenced the exit percentage, we calculated the conditional mean as a function of volume for each of five input quintiles. That is, for each site and for each input parameter, we classified each simulation into a quintile group based on the value for the considered parameter. Then, for each quintile group, we calculated a running mean conditional on the flow volume. These smoothed conditional means are depicted as curves in Figure 6. Within each domain, the difference between curves reflects the influence of the input parameter after we have controlled for volume.

High values of $m_0 - m_{\text{crit}}$ and log $k'$ were associated with smaller exit percentages. In contrast, variation in $\phi_{\text{bed}}$ had less influence. These results are similar to the influence of $m_0 - m_{\text{crit}}$ and log $k'$ in a study examining the 2014 Oso landslide (Iverson & George, 2016). The influence of the highest log $k'$ values (purple) is striking, explaining most of the simulations with low exit percentages. This result is consistent with the definition of log...
Most parameter combinations resulted in $V_e > 0$ m$^3$. Of the 400 simulations for each domain, 12, 54, and 4 simulations yielded $V_e = 0$ m$^3$ for Montecito, San Ysidro, and Romero domains, respectively. The San Ysidro domain was unlikely to have material exit the source basin at the lowest values of $V_e$. In contrast, the Romero domain produced high exit percentages across all volumes. This across-domain variation likely reflects the difference in the distribution of distance from outlet (Figures 5b and 5c). For the San Ysidro domain, a larger proportion of both the entire drainage area, as well as the area with initialized material, had large values for the distance from outlet. In contrast, those basins contributing to the Romero domain had relatively short distance to outlet.

### 7.2. Volume-Planimetric Area Scaling

Observations of debris flows, lahars, and rock avalanches follow an observed scaling relationship between volume, planimetric area, cross sectional area, and runout distance (Griswold & Iverson, 2008; Iverson et al., 1998; Rickenmann, 1999; Scheidl & Rickenmann, 2010). Empirical models such as LaharZ (Schilling, 1998) use these statistically estimated scaling relationships to forecast inundation area based on volume. One would expect that runout models considered here might reproduce the field observations relating volume and planimetric area. Such a comparison serves two distinct purposes. First, it is an important check on the models. Second, if successful, it may indicate that these models could be used to explore how flow properties and domain topography influence the volume area relationships.

For simplicity, we only considered planimetric area because the complex topography of the Montecito runout paths complicates comparisons of cross sectional area and runout distance. We considered multiple depth thresholds for identifying inundated area to reflect that the field observations for which these empirical relationships were calculated may not have included the thinnest deposits—such deposits may have been eroded for historical observations. We used $V_e$ for volume because it represents the debris flow volume which reached the inundated region and was most comparable to the debris flow deposits used to construct the empirical scaling relationship.

Observations of planimetric inundated area $B$ represent deposit extent, a quantity that was not directly simulated. As a proxy for deposit extent, we used the portion of the computational domain which saw flow depth greater than a specified threshold. There was not an obvious a priori threshold value to classify the peak flow depth extracted from simulations into inundated area; we considered 0, 0.1, 0.5, and 1.0 m. The value of $B$ was also limited at larger areas because we used a finite computational domain—when material flows out of this domain it did not contribute to $B$.

With depth thresholds <1 m, all models produced inundation extents that are larger than existing empirical relationships relating $B - V$ (Figure 7). In addition, the slope of the relationship between $B$ and $V_e$ is steeper than the 2/3 supported by field observations and the physical relationship between volume and area. Indeed, it is even steeper than the 0.81 identified by Bernard et al. (2021). The large threshold needed to come close to the observed relationship and the different slope likely reflect a combination of factors; these include (a) the preservation of large debris flow deposits, (b) the difference between areas that experienced transient inundation and deposition areas, (c) post-event redistribution of debris flow material, and (d) transport and deposition of suspended sediment by secondary fluvial processes.

In all models and all domains, the simulated inundation area matched the observations with a threshold of 0.5 m (Figure 7). For this reason, we used a threshold value of 0.5 m for classifying area as inundated and calculating the extent metric $\Omega_{in}$. This necessary choice impacts the subsequent results. However, it did not guarantee the overlap between simulated and observed inundation area because matching $B$ does not indicate the relative contribution of true positive, false positive, and false negative area. We recognize that this value is somewhat arbitrary and likely related to the inability of any model to simulate levee formation which steers the flow.

### 7.3. Interaction Between Outputs

Next we examined the interaction between each of the three elements of model performance that compose $C_r$ (Figure 8). A simulation with perfect performance would plot in the lower left of each panel at coordinates (0,0) and be colored yellow; no simulations plot at this coordinate. A simulation producing no inundation would score...
perfectly for excess inundation $\Delta_o$, because no model grid cell without observed inundation would be inundated (no false positive are produced). Such a simulation would yield the worst possible value for insufficient inundation $\Delta_u$, because no grid cells with observed inundation would be inundated (no true positive area produced). In contrast, a simulation that perfectly matched the extent of observed inundation but had a maximum depth of twice what was observed would score perfectly for insufficient inundation and a value of 1.0 for excess inundation.

Our sampling numerical design resulted in sets of simulations that demonstrate a trade off between not overinundating (low $\Delta_u$, high $\Delta_o$) and not underinundating (high $\Delta_u$, low $\Delta_o$) the simulation domain. Simulations that matched the extent well (low $\Omega_{\text{true}}$) were generally located at the lower-left apex of the $\Delta_o$-$\Delta_u$ curve, such that simulations that matched extent well had balanced performance in matching depth.

For some models and domains (e.g., Figure 8b), there is little variation along a tight $\Delta_o$-$\Delta_u$ trade-off curve, whereas in others there is considerable scatter away from this curve (e.g., Figure 8g). A tight curve reflects
that, conditional on the bounds of parameter space we have sampled, it was not possible to substantially change the extent of areas with excess inundation without also changing the areas that saw insufficient inundation. In contrast, when simulations fall off this curve, it reflects the possibility of changing volume or flow property parameters to result in changing performance in excess inundation without changing performance in insufficient inundation. For example, two simulations that inundate a similar area can do so with different average depths. This result exists in the RAMMS simulations (Figures 8a, 8d and 8g), and the observed behavior is explained by the $\mu_r$ value. Mobile material in simulations with high $\mu_r$ values stop flowing near the fan apex and spread out laterally, failing to inundate any portion of the lower fan.

7.4. Sensitivity to Volume

The relationship between each output and effective volume indicates variation across domains (Figure 9). As expected, simulations have lower values for $\Delta_o$ when volumes are low, and small values for $\Delta_u$ when volumes are high (Figures 9j–9o). This reflects difficulty in overestimating inundation depth with low volumes. For all outputs, there is scatter in model performance at any given volume value—this scatter reflects the variation in model performance from flow property parameter values (discussed in Section 7.5).

We might expect that the extent metric, $\Omega_{Tm}$, and the combination of $\Delta_o$ and $\Delta_u$ would have good performance for volumes that are close to the volume estimates for this event calculated in Section 2 (shown as vertical lines in Figure 9). These results were conditional on our choice of 0.5 m for the threshold for delineating inundated area. This threshold was chosen on the basis of total inundated area (true positive plus false positive), not on the basis of inundating the observed areas. All models show a “U-shaped” relationship between volume and these metrics, but only D-Claw has the trough of the U coincide with the expected volume. For both RAMMS and FLO-2D, the best performing simulations had volumes below the expected volume; though, as is discussed in Section 7.7, the performance of these two models did not substantially degrade when the estimated volume is used.
The absolute magnitude of $C_m$ for the best simulations varied by model and domain. All models had the most difficulty simulating the Romero domain and did the best on the San Ysidro domain. However, how the extent and depth metrics combine to create this pattern varied between model and domain. For example, D-Claw was able to simulate extent similarly for San Ysidro and Montecito domains, and depth similarly for San Ysidro and Romero domains. All models were able to simulate the depth best for the Montecito domain, owing to generally smaller values for $\Delta_o$. This was likely because this domain had deeper peak flow depths than the other two domains (Figure 1e).

The performance metrics for the best overall simulations for FLO-2D and D-Claw were similar, whereas RAMMS had slightly lower scores. However, the best scores are around 0.3, reflecting that all models are far from perfect at simulating the event. Only D-Claw was able to simulate the event at its best possible level in simulations with the $V_e$ in the range of volumes estimated for the event. Additionally, only D-Claw produced a minimum in the

---

**Figure 9.** Relationship between effective volume ($V_e$) and five metrics of model performance across model and domain. One metric is represented on each row and indicated by the row labels, each column indicates a different model, and each color reflects the domain. The y-axis limits are cropped such that values for $\Delta_\pi > 1$ and $(\Delta_s + \Delta_o)/2 > 1$ are not shown.
combined depth metric \((\Delta_p + \Delta_n)/2\) at the estimated event volume for all three domains. RAMMS and FLO2D did well at matching the depth for the Montecito domain, but their best performance for depth required smaller than reasonable volumes for the other two domains.

### 7.5. Sensitivity to Other Inputs

We identified the importance of input parameter values conditional on the simulated flow volumes. In this analysis, we only consider the combined metric \(C_m\). One approach to this analysis would be to look at the relationship between each input and \(C_m\); however, this examination would neglect the potentially confounding influence of volume, which we identified as important in Section 7.4. Instead, we considered the impact of varying each input, after controlling for the effective volume \(V_e\). For each combination of model, domain, and input parameter (each panel in Figure 10), we binned the simulations based on the input values into five equally spaced quintiles. Within each input parameter bin, we calculated the smoothed mean conditional on effective volume (curves in Figure 10). Within each bin, other input parameters are changing, resulting in variance around the curve. If \(C_m\) is insensitive to the considered input parameter, all of the bin curves will plot together. In contrast, shifts in the lines indicate the input parameter influences \(C_m\) after controlling for volume. Examining Figure 10 indicated that it is more common for an input to be unimportant than important. This result is certainly conditional on our choice of parameter ranges; we note our intention was to consider a wide, yet plausible range.

The most striking result of Figure 10 was how few non-volume inputs have a systematic influence on model results. For RAMMS, the parameter \(\mu_\text{r}\) is important at large values, with small \(\mu_\text{r}\) values having better performance. Small \(\mu_\text{r}\) values result in material that stops less easily, such that at large volumes, the flow does not stop and spreads out more. This result is observed for all three domains, most strongly for the San Ysidro domain. For FLO-2D, the only input that may indicate sensitivity is the slight performance improvement with smaller values of \(\mu_\text{r}\). D-Claw results show that the highest permeability values yielded poor results and that there may be a slight sensitivity to the basal friction angle. Recall that the volume considered here is the volume that leaves the basin, such that we have removed the influence of inputs on \(V_e\) identified in Section 7.1. Simulations with low permeability values resulted in smaller inundation areas with thicker flows, consistent with the definition of \(k'\).

### 7.6. Best Simulations

Our experimental design was not a formal calibration. Nevertheless, for each combination of site and model, it yielded simulations that were the best overall and the best conditional on using a volume (and \(m_\text{s}\) for D-Claw) comparable to the event estimate (Section 2). We examined these simulations to identify shared (or different) strengths and weaknesses across models and domains.

We compared the best overall simulations for each of the three model domains with each model (Figures 11–13). The results of Section 7.4 show the value of \(C_m\) for the best simulations for any given domain is similar across the three models. In addition, the locations where all models fail to match the simulated extent were similar (Figures 11–13). In the middle row of these figures, we indicate simulated material below the depth threshold of 0.5 m, and across all domains, both RAMMS and D-Claw produce large areas with thin inundation. This result is not surprising because none of the considered models represent the mechanisms of levee formation (e.g., Major & Iverson, 1999; Pudasaini & Fischer, 2020; Rocha et al., 2019). The lack of self-formed coarse-grained levees that serve to confine and steer the flow is a likely mechanism for the generation of widespread, thin flows.

Figures 11–13 provide visual context to the values of \(C_m\). Whereas all “best” simulations could be better in an absolute sense, all succeeded in capturing the pattern event. All models tended to create excess inundation and false positive area in topographically unconfined portions of the domain, especially those areas near the fan apex (that is, just downstream of the inflow locations in Figure 1b). Here, topographic confinement refers not just to the active stream channel but also to the meso-scale topography including fault-created hills and wider, less active alluvial fan channels. The most distal portions of Montecito Creek and San Ysidro domains were underinundated by all models. This may reflect the role of Highway 101 at the south end of the simulation domain. The highway provides both complex topography and a large depression. The inability of the models to simulate these areas points to the challenge of simulating the evolving solid volume fraction within a runout event. The actual inundation of this area was by thin, watery flows and was less hazardous than the areas proximal to the fan apex; thus, the inability to simulate this area well is less important for identifying high risk regions.
Figure 10. Model sensitivity to non-volume inputs. Each subplot reflects a model, input parameter, and domain. Each subplot title presents the model name (D = D-Claw; F = FLO-2D; R = RAMMS) and parameter symbol. The second line of the subplot title indicates the domain. Values of each input parameter were normalized to between 0–1 and binned based on quintiles (dot color). Curves reflect the mean conditional on volume for each input bin. Vertical dashed line and gray region indicate the factor of two uncertainty range on volume for each domain, respectively.
Figure 11. Comparison of best overall simulation in the Montecito domain for each model (columns). Top row shows the measured and modeled inundation extent, above a 0.5 m threshold, as elements of the binary classification confusion matrix (true negative omitted). Middle row shows the modeled peak flow depth. Bottom row shows the depth comparison residual as a function of measured depth with dots colored by the distance to the closest outlet.
Figure 12. Comparison of best overall simulation for the San Ysidro domain. See Figure 11 caption for explanation.
Many of the largest flow depth residuals were negative (bottom row in Figures 11–13) and from areas close to the fan apex (purple colors). This indicates that all models struggle to produce sufficiently high flow in the highest elevation (most upstream) portions of the fan. This is the case both for D-Claw, which directly simulates the flow hydrograph from initiation in the upland catchment, as well as for RAMMS and FLO-2D, which have a specified input hydrograph. Peak flow depth was estimated mainly from splash height, so the difficulty in matching peak flow depth may reflect both the nature of the data as well as the characteristics of the models.

The model input parameters for the best simulations are listed in Table 2. We exercise caution and focus interpretation on the parameters for which the models were sensitive (e.g., Figure 10). As we would expect from the sensitivity to volume, and sensitivity to other inputs results, the best simulations for RAMMS and FLO-2D had volumes that are smaller than the estimated volume for this event—an exception to this is the best RAMMS simulation for the Romero domain. Additionally, the best simulations all had low RAMMS $\mu_y$ values indicating flows that continue to run over low sloped regions of the domain.
Given the estimates of bulk solid volume fraction in Section 2 and the observations of $\frac{\bar{h}_s}{\bar{h}}$, we might expect that parameters for Montecito and Romero domains would be similar and also distinct from the San Ysidro domain. This is the case for the permeability results—though it is not clear that the results are sensitive to permeability so long as it is in the bottom 80%. The variation within this smaller range is likely unimportant given the spatial scale of the study.
7.7. Input Intercomparison

In contrast to the results illustrated in Figures 11–13 where we examined the best simulations for each domain and each model, our final numerical experiment focused on exchanging the best parameter values across domains. This isolated how model performance degrades when the best input values—conditional on using the correct volume and solid volume fraction—from one domain are used in another. Such an assessment starts to place constraints on the transferability of the results between watersheds, and potentially further afield. Additionally, we compared these simulations with ones chosen based on expert judgment.

Similar to the parameter sensitivity of Section 7.5, the most remarkable quality of this analysis (Figures 14–16) was how little variation there is between results obtained with different parameter sets and how little model results degraded when alternative parameter sets were used. Simulations with RAMMS and FLO-2D show more excess inundation across all three domains, reflecting these models’ inability to simulate the domains as well at the estimated volumes. These models also fail to inundate the most distal portions of Montecito and San Ysidro domains. D-Claw does not overinundate as much as the other two models—this is to be expected given that it performed its best at the correct volumes—but it also fails to inundate distal areas.

The relatively low variation in model performance that comes from this parameter set intercomparison is consistent with the earlier observation that few parameter values have a systematic influence on model performance (Figure 10). The expert opinion parameter sets did not substantially degrade model performance relative to the best-fit parameter set. The low level of simulation degradation under the parameter swap and expert opinion parameter sets supports the conclusion that getting the volume correct is the most important characteristic of simulating an event well. This may not seem surprising, yet it is useful to formally demonstrate. Other similar studies have come to opposite conclusions (e.g., Mergili et al., 2018); thus, we caution that this finding may not generalize to other locations, events, or choices regarding simulation performance assessment—for example, here we are concerned with inundation patterns and peak depths but not the details of timing. Additionally, in this event the flowing material was highly mobile.

8. Discussion

8.1. Relative Role of Volume and Flow Properties

By sampling across both volume and flow property inputs, our experimental design yielded the relative role of these two input categories. The results indicate that the primary control of simulation performance—as measured by $C_m$—was the volume. This implies that the most important pre-event uncertainty is how large an event will be. Given the wide variation in model governing equations, it is surprising that all three models were able to perform similarly and that so few of the input parameters showed systematic sensitivity. Whether this result persists at other locations or for other parameter ranges will require similar analysis of other events.

These results do not support the conclusion that the flow property inputs are completely unimportant. For example, Section 7.1 demonstrated that log $k'$ was influential in whether and how much material exits a source basin and inundates the fan. Further, we constrained D-Claw to consider only contractive ($m_0 - m_{crit} < 0$) flows because initially dilative material stops quickly. Identifying field observations that constrain these properties may reduce uncertainty in pre-event parameter selection. Additionally, there may be valuable insight relating these properties to initiation mechanisms—that is, do runoff-generated debris flows systematically have smaller $m_0$? Finally, we emphasize that it is only through further work that it will be possible to identify whether this relative unimportance of flow property inputs is a general finding or a characteristic of this location and event.

8.2. Pre-Event Uncertainty in Volume

Current capabilities to estimate pre-event volume rely on the empirical model of Gartner et al. (2014), which expresses sediment volume as a function of the $I_{15}$, basin relief, and the basin area burned at moderate to high severity. This empirical model is based on the sediment volume deposited into debris flow basins after individual events. Figure 3 shows the expected value and ±1 standard error range for sediment volume as a function of $I_{15}$ for each of the three considered domains. Owing to the large scatter in the original data, the range of the confidence interval is large—over an order of magnitude. The event estimates for the Montecito and Romero domains place the event near the lower bound of this range, and the San Ysidro domain is near the expected value.
Figure 14. Parameter value intercomparison for RAMMS. Each column uses the same material parameters, and each row uses the same domain. The flow property parameters for each column reflect the best simulation, conditional on volume, taken from each domain (Table 2). The inset box provides output metric values. Colors for true positive (dark green), false positive (light green), and false negative (light brown)—the same as in prior figures.
In a pre-event context, information about the peak 15-min rainfall intensity $I_{15}$ reflects meteorological forecasts and their associated uncertainty. Additionally, the total mobile volume reflects the combination of water and sediment. A full accounting for the pre-event uncertainty in flow volume must consider multiple uncertainty sources. One element is the forecast range in $I_{15}$ or historical precipitation-frequency data for estimates made in advance of a forecast timeframe. The uncertainty in $I_{15}$ then propagates into sediment volume. Incorporating plausible
variation in solid volume fraction then results in a range of total flow volumes. There are clearly physical links between soil characteristics, rainfall intensity and duration, and the mechanisms of debris flow initiation, which should be able to place further constraints on the volume of sediment and water mobilized in a given event.

The difference in volume initialization—in which RAMMS and FLO-2D used inflow hydrographs and D-Claw used material initialized upstream of the fan apex—motivates the following question: is it more feasible (or less
uncertain, or more practical) to know \( V_e \) or \( V_r \) before an event? The latter quantity focuses attention downstream of the fan apex, while the former additionally includes the processes by which material arrives at the fan apex. \( V_r \) is typically constrained with post-event field observations; for example, the Gartner et al. (2014) volume model yields values of \( V_r \) and is based on debris basin and sediment erosion records. Accordingly, our application of D-CLAW required evaluating the transfer function between \( V_r \) and \( V_e \) (Figure 6). Like all statistical relationships, however, extrapolation of the Gartner et al. (2014) volume model far beyond its training data may provide limited utility (e.g., Munroe, 2009). Estimation of \( V_r \) likely requires identification of initiation mechanisms and direct simulation of sediment mobilization (e.g., McGuire et al., 2016). While potentially more widely applicable, the practicality of such an approach has not been evaluated for hazard assessment purposes.

The comparison with existing area-volume scaling relationships (Figure 4) indicated that this event had a large inundation area for its volume, implying a relatively thin flow. In Figure 3, we showed that the event mobilized a relatively small quantity of sediment—relative to other events in the Gartner et al. (2014) compilation—given the rainfall intensity. These two observations together motivate further investigation of \( V-B \) relationships to understand the relative contribution of factors such as location characteristics (e.g., fan topography) and flow properties (e.g., solid volume fraction) in influencing the variance in inundated area.

### 8.3. Role of Basin and Fan Morphology

The best simulations for all models struggled in similar areas across the three considered domains. Most prominently, simulations produced overinundation in topographically unconfined areas and a lack of inundation in the most distal reaches of the domain. Overinundation was rare where topography confines the flow, even in areas proximal to the fan apex. For example, the moderate ridges in the center of the San Ysidro domain and the western bank of Montecito Creek both focus the flow direction and thus minimize inundation (stars indicated in Figure 1a).

These observations reflect the intrinsic characteristics of site morphology on the inundation hazard. Topographic characteristics alone may be sufficient to identify areas with low inundation hazard. These characteristics may be particularly important in topographically unconstrained fans—as compared with alpine or canyon debris flow fans, which are typically smaller and constrained by narrow valleys. The topography may provide general insight into areas of low and high inundation probability, but it does not reflect human structures which are known locations of increased avulsion potential. Studies at higher topographic resolution that compare inundation results with and without human structures may provide insight regarding their influence on inundation simulations.

Additionally, the analysis of basin exit percentage in D-Claw (Figure 6) demonstrated that source basin topography influenced how much mobilized material is able to reach the fan apex. Flow is highly constrained in the convergent channels of the upstream source basins, and the basin structure enforces the path a flow must take to reach the fan apex. The distance, slope, and set of tributary junction angles within a given basin represent knowable initial and boundary conditions for a hypothetical flow.

Considering the site characteristics and recalling the constraint of conservation of volume provides insight into the relative unimportance of flow properties identified in Section 7.5. Given a volume, the inundated area is controlled by the average deposit thickness. This quality underpins the success of scaling-type approaches (e.g., Iverson et al., 1998). The distribution of deposit depths is influenced by both flow properties and site topography.

### 8.4. Model Intercomparison and Natural Experiments

The governing equations described in Text S1 in Supporting Information S1 each represented a hypothesis of how the system dynamics may be simplified, and comparing the best simulation from each model provided a formal test of whether one approach simulated inundation patterns and peak depths better than the others (e.g., Barnhart, Tucker, et al., 2020a, 2020b). Overall D-Claw performed only slightly better than the other two models, especially at the estimated event volume (Table 2).

We found that \( C_m \) and its ingredients \( \Omega_m, \Delta_m, \) and \( \Delta_e \) have little information to distinguish between the three considered models; nevertheless, there may be other flow characteristics that contain distinguishing information (e.g., velocity, deposit depths, entrainment depths). However, to be widely usable, such flow characteristics must be measurable. A practical finding is that if inundation patterns are the primary quantity of interest, any model...
is sufficient, so long as the flow volume is known. There are many applications for which we anticipate a need for additional flow characteristics in order to assess risk (e.g., flow depth and velocity are relevant to estimate structure damage).

The 9 January 2018 Montecito debris flow event provides a natural experiment for assessing the relative role of volume, flow properties, and intrinsic site characteristics on inundation. Laboratory (or laboratory-like) experiments permit the necessary control for evaluating and advancing our understanding of physical processes, but natural experiments are a necessary window into assessing the influence of initial and boundary conditions at full scale. Evaluating the relative importance of these elements, relative to each other and physics is an exciting frontier for future work.

9. Conclusions

The core of this contribution is a model intercomparison of the 9 January 2018 Montecito debris flow inundation event. Based on comparing the flow volumes and inundated areas with existing data compilations, we find that the event was highly mobile, inundating a large area for its volume. Placing the estimated volume of sediment in context with empirical predictions for post-wildfire debris flows, we find that the event mobilized at or below the expected value of sediment, given the rainfall intensity.

We considered three different models (RAMMS, FLO-2D, and D-Claw) and find that—so long as flow volume is constrained well—all three did well at simulating the inundation patterns and peak depths observed in the event. The parameters that control flow properties were relatively unimportant, though this result may depend on the specific parameter range considered. The similar success of all models and the low sensitivity to input parameters may be associated with the high flow mobility. The result that flow volume is most important is not surprising, but it points to the importance of constraining a difficult quantity to predict. Indeed, the most successful tools for predicting post-wildfire debris flow volumes in a pre-event context carry order of magnitude uncertainties. Forensic back-calculations can rely on estimates of flow volume; in contrast, pre-event hazard assessments must use forecasts of event volume.

We compared three source basins and runout domains and found that both fan and basin morphology influenced model performance, assessed based on matching observed inundation patterns and peak flow depths. The influence of basin and fan morphology points to the role of initial and boundary conditions on the evolution of the flow dynamics. The role of the runout domain topography may partially explain why all models were equally successful at simulating the event.

Because we used models with different initialization schemes, we were able to compare simulated hydrographs with ones based on the Rickenmann (1999) scaling relationships and show that the simple triangular hydrographs capture the simulated dynamics well. Whereas FLO-2D and RAMMS expect a user to know the shape of an input hydrograph to initialize the model, D-Claw was developed with different expectations for what must be known for initialization. This meant we could not guarantee a specific flow volume leaving the basin in our D-Claw simulations, and it supported our analysis showing that the $k'$ and $m_0 - m_{crit}$ input parameters influenced the volume of material that exited the source basins.

This study compared models with data; we compared simulated and observed inundation presence and peak flow depth to construct a performance metric that differentiated between relatively good and bad model performance. Because no performance metrics are established for evaluating inundation simulations, we recognize that our results are conditional on the choice of metric. Another metric that incorporated additional observations might produce a different result. Flow velocity (either within the flow or of the flow front) may better evaluate event dynamics. Additionally, evaluation of inundation presence may be refined (e.g., based on the meso-scale topographic context). Any broadly usable metric must be based on observations that are possible to collect in the field.

Data Availability Statement

Data were not created for this research. The results presented here used three model codes: RAMMS, FLO-2D, and D-Claw. RAMMS was originally described by Christen et al. (2010). It is proprietary software and more information about it can be found at https://ramms.sil.ch/ramms/. The core of the FLO2D shear stress versus strain
rate relationship was developed by O’Brien and Julien (1985), and the most comprehensive description of it can be found in the FLO-2D user manual (O’Brien, 2020a). It is also proprietary software, and more information about it can be found at https://fло-2d.com/. D-Claw is an open source model originally described by George and Iverson (2014) and Iverson and George (2014). Its source code is available on GitHub at https://github.com/geoflows/D-Claw.

Acknowledgments
We thank Editor Amy East, Associate Editor Odin Marc, three anonymous reviewers, Dick Iverson, Jeff Coe, Brian Shiro, and Janet Slate for constructive comments that substantially improved the quality and clarity of the manuscript. This research was funded by the U.S. Geological Survey Landslide Hazard Program. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Adams, B., Bauman, L., Bohnhoff, W., Dalbey, K., Ebeida, M., Eddy, J., & Winokur, J. (2020). Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 6.12 user manual. Sandia Technical Report SAND2020-5001.

Barnhart, K. R., Hutton, E. W. H., Tucker, G. E., Gasparini, N. M., Istanbulbullouglu, E., Hobley, D. E. J., et al. (2020). Short communication: Landlab v2.0: A software package for Earth surface dynamics. Earth Surface Dynamics, 8(2), 379–397. https://doi.org/10.5194/esurf-8-379-2020

Barnhart, K. R., Tucker, G. E., Doty, S., Shobe, C. M., Glade, R. C., Rossi, M. W., & Hill, M. C. (2020a). Inverting topography for landscape evolution model process representation: Part 1, Conceptualization and sensitivity analysis. Journal of Geophysical Research: Earth Surface, 125(7), e2018JF004961. https://doi.org/10.1029/2018JF004961

Barnhart, K. R., Tucker, G. E., Doty, S., Shobe, C. M., Glade, R. C., Rossi, M. W., & Hill, M. C. (2020b). Inverting topography for landscape evolution model process representation: Part 2, Calibration and validation. Journal of Geophysical Research: Earth Surface, 125(7), e2018JF004963. https://doi.org/10.1029/2018JF004963

Bartelt, P., Bieler, C., Bühler, Y., Christen, M., Deubelbeiss, Y., Graf, C., & Schneider, M. (2021). RAMMS::DERRISFLOW user manual v1.7.0. WSL Institute for Snow and Avalanche Research SLF. Retrieved from https://ramms.slf.ch/ramms/downloads/RAMMS_DBF_Manual.pdf

Bernard, D., Troulis, E., & Santi, P. (2021). Estimation of inundation areas of post-wildfire debris flows in southern California USA. Engineering Geology, 285, 105991. https://doi.org/10.1016/j.enggeo.2021.105991

Bertolli, P., & Wieczorek, G. F. (2005). Calibration of numerical models for small debris flows in Yosemite Valley, California, USA. Natural Hazards and Earth System Sciences, 5(6), 993–1001. https://doi.org/10.5194/nhess-5-993-2005

Bessez-Norton, E., Kien, J. W., Coe, J. A., Rengers, F. K., & Staley, D. M. (2019). An evaluation of debris-flow runout model accuracy and complexity in Montecito, CA: Towards a framework for regional inundation-hazard forecasting. Seventh International Conference on Debris-Flow Hazards Mitigation. https://doi.org/10.25567/11124/173211

CAL FIRE. (2018). Thomas fire watershed emergency response team final report (Technical Report). Retrieved March 17, 2021, from http://cdfdata.fire.ca.gov/admin8327985/cdf/images/incidentfile1922_3383.pdf

Cannon, S. H., Garnier, J., Rupert, M., Michael, J., Staley, D., & Worstell, B. (2009). Emergency assessment of postfire debris-flow hazards for the 2009 Station Fires, San Gabriel Mountains, Southern California (p. 27). U.S. Geological Survey Open-File Report 2009-1227. https://doi.org/10.1029/20091227

Cannon, S. H., Kirkham, R. M., & Parise, M. (2001). Wildfire-related debris-flow initiation processes, Storm King Mountain, Colorado. Geomorphology, 39(3), 171–188. https://doi.org/10.1016/S0169-555X(00)00108-2

Christen, M., Kowalski, J., & Bartelt, P. (2010). RAMMS: Numerical simulation of dense snow avalanches in three-dimensional terrain. Cold Regions Science and Technology, 63(1), 1–14. https://doi.org/10.1016/j.coldregions.2010.04.005

Chu, S. T. (1978). Infiltration during an unstable rain. Water Resources Research, 14(3), 461–466. https://doi.org/10.1029/WR014i003p00461

Di Santolo, A. S., & Evangelista, A. (2009). Some observations on the prediction of the dynamic parameters of debris flows in pyroclastic deposits in the Campania region of Italy. Natural Hazards, 50(3), 605–622. https://doi.org/10.1007/s11069-008-9334-3

Fan, L., Lehmann, P., Mcardell, B., & Or, D. (2017). Linking rainfall-induced landslides with debris flows runout patterns towards catchment scale hazard assessment. Geomorphology, 280, 1–15. https://doi.org/10.1016/j.geomorph.2016.10.007

Frank, F., Mcardell, B. W., Oggier, N., Baer, P., Christen, M., & Vieli, A. (2017). Debris-flow modeling at Meretschibach and Bondasca catchments, Switzerland. Sensitivity testing of field-data-based entainment model. Natural Hazards and Earth System Sciences, 17(5), 801–815. https://doi.org/10.5194/nhess-17-801-2017

Gartner, J. E., Cannon, S. H., & Santi, P. M. (2014). Empirical models for predicting volumes of sediment deposited by debris flows and sediment-laden floods in the transverse ranges of southern California. Engineering Geology, 176, 45–56. https://doi.org/10.1016/j.enggeo.2014.04.008

George, D. L., & Iverson, R. M. (2014). A depth-averaged debris-flow model that includes the effects of evolving dilatancy: II. Numerical predictions and experimental tests. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 470(2170), 20130820. https://doi.org/10.1098/rspa.2013.0820

Grissvold, J. P., & Iverson, R. M. (2008). Mobility statistics and automated hazard mapping for debris flows and rock avalanches (p. 59). U.S. Geological Survey Scientific Investigations Reports 2007-5276 (Version 1.1, April 2014). https://doi.org/10.3133/si20075276

Heiser, M., Scheidl, C., & Kaitna, R. (2017). Evaluation concepts to compare observed and simulated deposition areas of mass movements. Computational Geosciences, 21(3), 335–343. https://doi.org/10.1007/s10596-016-9609-9

Hobley, D. E. J., Adams, J. M., Nudurupati, S. S., Hutton, E. W. H., Gasparini, N. M., Istanbulbullouglu, E., & Tucker, G. E. (2017). Creating computing with Landlab: An open-source toolkit for building, coupling, and exploring two-dimensional numerical models of Earth-surface dynamics. Earth Surface Dynamics, 5, 21–46. https://doi.org/10.5194/esurf-5-21-2017

Hung, O., & McDougall, S. (2009). Two numerical models for landslide dynamic analysis. Computers and Geosciences, 35(5), 978–992. https://doi.org/10.1016/j.cageo.2007.12.003

Hussein, H. Y., Quan Luna, B., van Westen, C. J., Christen, M., Malet, J. P., & van Asch, T. W. J. (2012). Parameterization of a numerical 2D debris flow model with entainment: A case study of the Faucon catchment, southern French Alps. Natural Hazards and Earth System Sciences, 12(10), 3075–3090. https://doi.org/10.5194/nhess-12-3075-2012

Ice, G. G., Neary, D. G., & Adams, P. W. (2004). Effects of wildfire on soils and watershed processes. Journal of Forestry, 102(6), 16–20.

Iverson, R. M. (1997). The physics of debris flows. Reviews of Geophysics, 35(3), 245–296. https://doi.org/10.1029/97RG00426

Iverson, R. M. (2003). The debris-flow rheology myth. 3rd International Conference on Debris-Flow Hazards Mitigation: Mechanics, Prediction, and Assessment (pp. 303–314).
Rocha, F. M., Johnson, C. G., & Gray, J. M. N. T. (2019). Self-channelisation and levee formation in monodisperse granular flows. *Journal of Fluid Mechanics, 876*, 591–641. https://doi.org/10.1017/jfm.2019.518

Salm, B. (1993). Flow, flow transition and runout distances of flowing avalanches. *Annals of Glaciology, 18*, 221–226. https://doi.org/10.3189/026030593781868614

Salm, B., Burkard, A., & Gabhler, H. U. (1990). *Berechnung von Fließlawinen: Eine Anleitung fuer Praktiker mit Beispielen*. Eidgenössische Institut für Schnee- und Lawinenforschung.

Santi, P. M., & Rengers, F. K. (2020). Wildfire and landscape change. In *Reference module in earth systems and environmental sciences*. Elsevier. https://doi.org/10.1016/B978-0-12-818234-5.00017-1

Scheidl, C., & Rickenmann, D. (2010). Empirical prediction of debris-flow mobility and deposition on fans. *Earth Surface Processes and Landforms, 35*(2), 157–173. https://doi.org/10.1002/esp.1897

Schilling, S. P. (1998). *LAHARZ; GIS programs for automated mapping of lahar-inundation hazard zones* (Technical Report). Retrieved from https://pubs.er.usgs.gov/publication/ofr98638

Schraml, K., Thomschitz, B., Mc Ardell, B. W., Graf, C., & Kainra, R. (2015). Modeling debris-flow runout patterns on two alpine fans with different dynamic simulation models. *Natural Hazards and Earth System Sciences, 15*(7), 1483–1492. https://doi.org/10.5194/nhess-15-1483-2015

Staley, D. M., Negri, J. A., Kean, J. W., Laber, J. L., Tillery, A. C., & Youberg, A. M. (2017). Prediction of spatially explicit rainfall intensity-duration thresholds for post-fire debris-flow generation in the western United States. *Geomorphology, 278*, 149–162. https://doi.org/10.1016/j.geomorph.2016.10.019

Staley, D. M., Negri, J. A., Kean, J. W., Tillery, A. C., & Youberg, A. M. (2016). *Updated logistic regression equations for the calculation of post-fire debris-flow likelihood in the western United States* (p. 13). U.S. Geological Survey Open-File Report. https://doi.org/10.3133/ofr20161106

Swiler, L. P., & Wyss, G. D. (2004). A user’s guide to sandia’s Latin hypercube sampling software: LHS UNIX library/standalone version. Sandia Technical Report. https://doi.org/10.2172/919175

U. S. Forest Service. (2020). *Thomas burned area report, FSH 2509.13* (Technical Report). U.S. Department of Agriculture. Retrieved from https://www.fs.usda.gov/detail/lpnf/home/?cid=fseprd570093

U.S. Geological Survey. (2017). *Lidar point cloud - USGS national map 3DEP downloadable data collection: U.S. Geological Survey*. Retrieved from https://www.sciencebase.gov/catalog/item/470ab664e4b058caae3f0def

U.S. Geological Survey. (2018). *Emergency assessment of post-fire debris flow hazards: Thomas fire* (Technical Report). Retrieved from https://landslides.usgs.gov/hazards/postfire_debrisflow/

U.S. Geological Survey Advanced Research Computing. (2021). *USGS Denali Supercomputer: U.S. Geological Survey*. https://doi.org/10.5066/P9PSW367

Wehrli, S. (2019). *The influence of digital elevation model data on debris flow runout modelling*. (Master’s thesis, University of Zurich). Retrieved from https://lean-gate.geo.uzh.ch/prod/index.php?id=mscthesispdf&maid=670

References From the Supporting Information

Berger, M. J., George, D. L., LeVeque, R. J., & Mandli, K. T. (2011). The GeoClaw software for depth-averaged flows with adaptive refinement. *Advances in Water Resources, 34*(9), 1195–1206. https://doi.org/10.1016/j.advwatres.2011.02.016

Clawpack Development Team. (2019). *Clawpack software*. https://doi.org/10.1016/j.advwatres.2011.02.016

Daerr, A., & Douady, S. (1999). Two types of avalanche behaviour in granular media. *Nature, 399*(6733), 241–243. https://doi.org/10.1038/20392

Mandli, K. T., Ahmadia, A. J., Berger, M., Calhoun, D., George, D. L., Hadjimichael, Y., & LeVeque, R. J. (2016). *Clawpack: Building an open source ecosystem for solving hyperbolic PDEs*. *PeerJ Computer Science, 2*, e68. https://doi.org/10.1038/20392