Hydrological Processes

A Multi-sensor Evaluation of Precipitation Uncertainty for Landslide-triggering Storm Events

Journal:	*Hydrological Processes*
Manuscript ID	Draft
Wiley - Manuscript type:	Special Issue Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Culler, Elsa; University of Colorado Boulder, Civil, Environmnetal, and Architectural Engineering; University of Colorado Boulder, Collaborative Institute for Research in Environmental Science (CIRES) Badger, Andrew; USRA, GESTAR; NASA Goddard Space Flight Center, Minear, J.; University of Colorado Boulder, Collaborative Institute for Research in Environmental Science (CIRES) Tiampo, Kristy; University of Colorado Boulder, Collaborative Institute for Research in Environmental Science (CIRES); University of Colorado Boulder, Department of Geological Sciences Zeigler, Spencer; University of Colorado Boulder, Department of Geological Sciences Livneh, Ben; University of Colorado Boulder, Department of Civil, Environmental, and Architectural Engineering; University of Colorado Boulder, Collaborative Institute for Research in Environmental Science (CIRES)
Keywords:	precipitation inter-comparison, rainfall-triggered landslides, natural hazards, extreme precipitation, intensity-duration thresholds

SCHOLARONE™
Manuscripts

A Multi-sensor Evaluation of Precipitation Uncertainty for Landslide-triggering Storm Events

Landslide Precipitation Inter-comparison

Elsa S. Culler [1,2], Andrew M. Badger [3,4], J. Toby Minear [2], Kristy F. Tiampo [2,5], Spencer D. Zeigler [5], and Ben Livneh [1,2]

[1] Department of Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, Boulder, CO 80309, USA

[2] Cooperative Institute for Research in Environmental Science (CIRES), University of Colorado Boulder, Boulder, CO 80309, USA

[3] Universities Space Research Association, Columbia, MD, 21046 USA

[4] Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, 20771 USA

[5] Department of Geological Sciences, University of Colorado Boulder, Boulder, CO 80309, USA

Corresponding Author: Elsa Culler, 216 UCB, University of Colorado Boulder campus, Boulder, CO 80309, elsa.culler@colorado.edu

Acknowledgements

This research was supported by funding from the National Aeronautics and Space Administration's Interdisciplinary Research in Earth Science (IDS) program grant 16-IDS16-0075, The Interaction of Mass Movements with Natural Hazards Under Changing Hydrologic Conditions.

Conflict of Interest Statement

The authors declare no conflict of interest

For Peer Review

Abstract: Extreme precipitation can have profound consequences for communities, resulting in natural hazards such as rainfall-triggered landslides that cause casualties and extensive property damage. A key challenge to understanding and predicting rainfall-triggered landslides comes from observational uncertainties in the depth and intensity of precipitation preceding the event. Practitioners and researchers must select among a wide range of precipitation products, often with little guidance. Here we evaluate the degree of precipitation uncertainty across multiple precipitation products for a large set of landslide-triggering storm events and investigate the impact of these uncertainties on predicted landslide probability using published intensity-duration thresholds. The average intensity, peak intensity, duration, and NOAA-Atlas return periods are compared ahead of 228 reported landslides across the continental US and Canada. Precipitation data are taken from four products that cover disparate measurement methods: near real-time and post-processed satellite (IMERG), radar (MRMS), and gauge-based (NLDAS-2). Landslide-triggering precipitation was found to vary widely across precipitation products with the depth of individual storm events diverging by as much as 296 mm with an average range of 51 mm. Peak intensity measurements, which are typically influential in triggering landslides, were also highly variable with an average range of 7.8 mm/hr and as much as 57 mm/hr. The two products more reliant upon ground-based observations (MRMS and NLDAS-2) performed better at identifying landslides according to published intensity-duration storm thresholds, but all products exhibited hit-ratios of greater than 0.56. A greater proportion of landslides were predicted when including only manually-verified landslide locations. We recommend practitioners consider low-latency products like MRMS for investigating landslides, given their near-real time data availability and good performance in detecting landslides. Practitioners would be well-served considering more than one product as a way to confirm intense storm signals and minimize the influence of noise and false alarms.

Keywords: precipitation inter-comparison, rainfall-triggered landslides, natural hazards, extreme precipitation, intensity-duration thresholds

1 INTRODUCTION

Precipitation measurements and their uncertainties play a key role in understanding and mitigating rainfall-triggered landslides because they drive excess runoff and soil saturation that initiate these natural disasters (Highland & Bobrowsky, 2008). In spite of the destructive nature of landslides, which cause tens of thousands of deaths each year (Froude & Petley, 2018) these events remain challenging to diagnose, in part due to uncertainty in antecedent precipitation amounts (Kirschbaum & Stanley, 2018). There are many other sources of uncertainty that contribute to poor landslide diagnosis and prediction, such as unknown soil properties, vegetation, and anthropogenic modifications to surface and subsurface soil structure. However, perhaps the largest source of uncertainty in estimating landslide probability is hydrologic uncertainty, defined here as uncertainty in

For Peer Review

the depth and intensity of liquid precipitation leading up to the event (Chowdhury & Flentje, 2002). A confounding factor is the wide array of precipitation datasets ranging from in situ observations, ground-based radar and satellite retrievals. The goals of this study are to investigate the role of precipitation uncertainty preceding known historical landslide events, and to assess the implications for evaluating landslide hazards.

The precipitation products chosen for this inter-comparison represent three broad categories of primary measurement techniques: precipitation gauges, ground-based radar, and microwave satellite. Precipitation gauges operate by periodically measuring the volume of precipitation collected at a gauge. Their main strength is that they directly measure the amount of collected water, but nonetheless they can suffer from issues of persistent bias driven by under-catch from wind (Pollock et al., 2018) instrument malfunctions (Duchon et al., 2014; Duchon & Biddle, 2010), gauges placed too close to other structures (Vose et al., 2014), and limited spatial representativeness due to sparse sensor density (Kidd et al., 2017).

In contrast, ground-based radar detects precipitation indirectly using the backscatter of radar and can measure subtle variations in precipitation over regions of several hundreds of square kilometers (Zhang et al., 2015). Since ground-based radar is an indirect measurement of precipitation, its performance is dependent on skilful conversion of the radar signal to precipitation volume. Beam blockage and interference from buildings or even insects in the radar's path are another limitation (Bousquet & Smull, 2003; Fornasiero et al., 2004; Nikahd et al., 2016). Most ground-based radars use multiple bands of radar and multiple polarities to compute the raindrop shape and size distributions used in the processing and limit the impact of known sources of error, which offers an advantage over other indirect techniques such as some of those incorporated into satellite-based measurements (Chandrasekar et al., 2008).

Satellite techniques vary in terms of which sensors they use to detect precipitation, including active and passive microwave, infrared, radar, or any combination. Depending on the sensor type these satellites can be deployed in either geostationary or low Earth orbits that cover particular spatial regions at particular intervals (Huffman et al., 2020). The key advantage of satellite-based precipitation measurements over ground-based in situ or radar sensors is that they can deliver frequent and spatially continuous measurements, although multiple satellites (Tapiador et al., 2012) with a variety of sensors and orbits (Ashouri et al., 2015) are required to provide global coverage. For example, the satellite products used in this analysis incorporate a fleet of geostationary satellites in addition to a single low Earth orbit reference satellite (Kidd et al., 2020). Many of the challenges associated with satellite-based precipitation measurement are related to sensor calibration and bias-correction relative to ground-based measurements (Ebert, 2007), as well as the development of algorithms for merging measurements from diverse sources (Huffman et al., 2007; Skofronick-

For Peer Review

Jackson et al., 2017). Estimating drop size distributions is also a challenge, though it can be addressed through the use of either ground- or satellite-based radar.

Existing precipitation products have been compared and evaluated using a number of metrics in prior studies, for example annual and monthly totals (Adler et al., 2001) or the frequency of wet or dry days (Manzanas et al., 2014). Less attention has been paid to metrics most directly useful for analysing rainfall-triggered landslides. While some landslides are triggered by short, intense precipitation events, others are triggered by saturation of the soil column that can develop over a longer period of time (Cannon & Gartner, 2005). In both cases the triggering event occurs over the course of hours or days rather than months or years, and for some landslides the critical time period may be less than an hour.

Published precipitation inter-comparisons typically focus on specific applications such as evaluating grid-based products over complex terrain, portraying hydrologic phenomena (Ahmadalipour & Moradkhani, 2017), climate model downscaling efforts (Gutmann et al., 2014; Wang et al., 2020), or for merging multiple sensors together (Beck et al., 2017). A general review of 30 gauge-based, satellite-based, and reanalysis global precipitation products by Sun et al. (2018) compared systematic and random errors for daily and annual precipitation, reporting large disagreements even within the same class of product, i.e., a deviation of 300 mm in annual precipitation over global land among satellite products. They conclude that the placement and density of gauges accounts for many of the errors in gauge-based or gauge-corrected products, further suggesting that cross validation across multiple datasets is crucial to account for errors. Adler et al. (2003) similarly analyzed 31 gauge-based, satellite-based, model-based, and climatological datasets in terms of monthly precipitation, finding that 'quasi-standard' products, e.g., those like the Global Precipitation Measurement (GPM) mission (Hou et al., 2014) that have undergone substantial testing, perform better. Additionally, they report that products incorporating both in situ and satellite information (e.g., the Global Precipitation Climatology Project) perform better than products based on a single data source.

Fewer studies comparing extreme precipitation events (e.g., events above the 90th percentile) exist. Many focus on climate model simulations (Sunyer et al., 2015; Tryhorn & DeGaetano, 2011) and trends (Bao et al., 2017; Janssen et al., 2014) while others evaluate observations and satellites (AghaKouchak et al., 2011; Lockhoff et al., 2014; Pendergrass & Knutti, 2018). AghaKouchak et al. (2011) compared extreme precipitation across four satellite platforms and found trade-offs across products in terms of correct identification of precipitation above a threshold and measurements of the volume of extreme storms. While they showed that some datasets performed better than others in certain contexts, they ultimately concluded that no single precipitation product was ideal for detecting extremes because all of them failed to detect certain storms in certain regions. Lockhoff et al. (2014) found that satellite retrieved extreme precipitation values generally matched station-based

For Peer Review

precipitation when using fuzzy metrics to evaluate agreement at larger spatiotemporal scales of ~330 km and 5 days. Pendergrass & Knutti (2018) showed that precipitation was less variable in coarser versus finer-resolution satellite precipitation datasets, suggesting that coarser precipitation products may be unable to capture extreme precipitation to the same extent as higher resolution datasets. Other studies primarily evaluated extreme precipitation indicators like 90th percentile precipitation, extreme one-day precipitation and maximum number of consecutive wet days (Amitai et al., 2012; Manzanas et al., 2014). These measures are meant to capture large storms that happen on at least an annual basis rather than specific to storms that trigger natural disasters (Manzanas et al., 2014; Sun et al., 2018).

This works builds on the handful of studies that have specifically evaluated multiple precipitation products in the context of landslide triggering (Rossi et al., 2017; Brunetti et al., 2018; Tajudin et al., 2020; Chikalamo et al., 2020). For example, Rossi et al. (2017) compared of satellite and gauge precipitation data preceding landslide events in Italy, using intensity-duration thresholds as a part of the comparison. They found that data from Tropical Rainfall Measuring Mission (TRMM) satellite products (Kummerow et al., 1998) tend to underestimate gauge data, particularly in mountainous areas where landslides are most likely to occur. Brunetti et al. (2018) similarly found that satellite precipitation from four products tended to underestimate rainfall relative to ground based observations. Both studies ultimately concluded that the satellite data could still be useful for forecasting landslides as long as issues of local bias could be accounted for.

The intensity-duration threshold is a type of two-parameter statistical model used for landslide early warning systems, where rainstorms above the threshold curve are predicted to cause landslides (Scheevel et al., 2017). The curves are typically based on a power law (e.g., $I = aD^{-b}$) of storm intensity (I) as a function of duration (D) with fitted parameters a and b. These power laws are valid in a particular region or climate and for a range of durations depending on the training data (Guzzetti et al., 2008). Other statistical rainfall thresholds have been proposed, but generally rely upon either intensity or duration or both (Galanti et al., 2018; Leonarduzzi et al., 2017). Here, we will investigate several power-law intensity-duration thresholds reviewed by Guzzetti et al. (2008) as a basic way to compare precipitation measurements from different sources in the context of landslide hazard estimation. Furthermore, when precipitation is used to provide warning systems or guide recovery efforts from landslides, the timeliness, i.e., low latency, of the information matters (Kirschbaum et al., 2012), such that the issue of latency will also be considered in the investigation of intensity-duration thresholds.

The focus of this analysis is to quantify precipitation uncertainty associated with known historical landslides, and to examine the role of this uncertainty in modelling landslide hazards. Given the wide-ranging issues associated with precipitation estimation cited above, this study presents a multi-product, multi-site analysis focused on landslide-triggering storms. We address an existing gap in

For Peer Review

evaluating extreme precipitation through the lens of rainfall-triggered landslide hazards, while conducting inter-product analyses into storm characteristics of potential relevance for the hydrological community. Additionally, we further the analyses by Rossi et al. (2017) and Brunetti et al. (2018) by including ground-based radar and by rigorously analyzing each precipitation estimate preceding specific landslide events. Greater understanding of the areas of relative agreement and any divergence across products may provide guidance to practitioners and researchers choosing among precipitation products for studying landslides.

2 METHODS

We compared precipitation characteristics at known landslide sites across the features of triggering storms, as well as relative to intensity-duration thresholds of landslide occurrence. Rainfall-triggered landslide sites were chosen from the NASA Global Landslide Catalog (GLC; Kirschbaum et al., 2010) with a subset of landslide locations verified with ancillary satellite imagery (see section 2.1). For each landslide location, precipitation was obtained from four different products (see section 2.2) and the precipitation time series were split into individual storm events. For each storm, key characteristics of total depth, duration, intensity, peak intensity, and return period were calculated (section 2.3). Finally, the storm events were plotted relative to landslide intensity-duration curves, with hit-ratios and false-alarm-ratios compared for each model-product combination (section 2.4).

2.1 Study domain and landslide site selection

The NASA Global Landslide Catalog (GLC; Kirschbaum et al., 2010) was chosen as the source of landslide locations for this study, since it provides a large sample of landslide locations useful for evaluating heavy rainfall events. The GLC shares many strengths and weaknesses with other regional and global databases available (Kirschbaum et al., 2010; Mirus et al., 2020). Though the GLC covers a broad spatial and temporal domain, it suffers from problems of precision and completeness. The catalog is comprised of a collection of second-hand landslide reports made by organizations like the news media, governmental organizations like departments of transportation, along with available scientific reports. This means that landslides that nearby infrastructure and people are reported more frequently, resulting in a substantial spatial bias towards populated areas. Landslide location accuracies range from ‘exact’ locations, to location uncertainties between 1km up to 50 km, depending on how specific the source article was about the location (Kirschbaum et al., 2010). Despite these limitations, the GLC was deemed fit for the purposes of this study, which is not to study landslide mechanisms and spatial distribution, but rather to compare precipitation products in the vicinity of hydrologically-triggered landslides where heavy rainfall events are likely to be present. Overall, the GLC provided a substantial number of landslide locations (n=228) for this study that met the following selection criteria:

For Peer Review

- Only landslide events reported as rainfall-driven, with a GLC trigger category of "rain," "downpour," "continuous_rain," or "flooding" were included. Snow-related triggers were not included even though these are hydrologically driven, because their precipitation is not contemporaneously linked with landslide triggers;
- Landslide events took place in the continental United States (CONUS) or Canada below 60° N and after May 2015 ensuring data availability across each of the selected precipitation products;
- The landslide location accuracy was reported to be 10 km or less. The value of 10 km was chosen since it is approximately equal to the spatial resolution of two of the precipitation products; and
- The landslide size was reported as "medium" or larger so as to select for events more likely to have been triggered by substantial precipitation.

In total, 228 landslides were selected. Of those, the exact locations for 80 sites were verified by a trained technician searching for a landslide scarp in visible satellite images of the terrain near the specified landslide location. The location specified by the GLC was used for the remaining landslides where 31 were marked in the GLC as "exact" locations, 51 as 1 km, 52 as 5 km, and 14 as 10 km accuracy. Figure 1 shows that many of the sites are located near the Pacific coast, likely due to the presence of complex topography associated with landslides, as well as the population reporting bias of the GLC. The verified landslides are generally distributed evenly relative to the locations of the full selection of landslides.

[Insert Figure 1]

2.2 Precipitation data sources

The gridded precipitation datasets in this study were chosen to be reflective of three common measurement methods: gauges, ground-based radar, and satellite. We were interested in products that are freely available, have undergone extensive verification, and extend over at least the CONUS. An important additional criterion was that products be available at an hourly temporal resolution or finer in order to compute the characteristics of individual storm events. We further sought to include products with multiple latencies where available. The above criteria resulted in the precipitation products and features described in Table 1 and summarized below.

[Insert Table 1]

North American Land Data Assimilation System version 2 (NLDAS-2) meteorological dataset

For Peer Review

The NLDAS-2 meteorological dataset (Xia et al., 2012) is a combination of daily gauge-based National Center for Environmental Prediction (NCEP) Climate Prediction Center (CPC) precipitation with orographic corrections and hourly NCEP Doppler radar-based precipitation. The gauge-based estimates are disaggregated to hourly using the radar-based estimates, resulting in a near real-time hourly gridded product at 0.125° (~12 km) resolution across North America going back to 1979 with a latency of approximately four days. Though it has coarser horizontal resolution relative to the other precipitation products used here, NLDAS-2 meteorological is a widely used gauge-based product that has been extensively validated over a recent period overlapping with this study (Livneh et al., 2015; Long et al., 2014; Xia et al., 2016).

Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimate

MRMS precipitation estimates are primarily based on a centralized radar mosaic with 2-minute resolution over the US and Canada. This study uses an hourly version that also integrates data from numerical weather prediction, satellites, gauges, lightning sensors, and precipitation models (Zhang et al., 2015). While both NLDAS-2 and MRMS estimates contain common information from gauges and radar, the NLDAS-2 product is primarily a gauge-based estimate while MRMS focuses on radar inputs. MRMS is the precipitation product with the shortest period of record among the products selected for this study, and so there are relatively few years of data for validation. However, it has by far the highest resolution at 0.01° (~1.1 km) and represents the state of the art in terms of leveraging computing resources to take advantage of a multitude of overlapping radar and other types of sensors.

Integrated Multi-satellitE Retrievals for GPM (IMERG)

GPM IMERG precipitation estimates are a combination of multiple satellite measurements, including the GPM Core Observatory Microwave Imager which is considered the standard for the other included satellites. In addition to active and passive microwave sensors, IMERG estimates include Infrared sensors, satellite-based radar, and precipitation gauge adjustments. The gauges are used for monthly bias correction (Huffman et al., 2020). There are three IMERG products, Early, Late, and Final, of which we use the Early (~4-hour latency) and the Final (~3.5-month latency) in this study. The IMERG-Early product is available much more promptly than the IMERG-Final, but as a result some of the satellite retrievals are not incorporated because they have not yet arrived, and it cannot take advantage of some processing steps or monthly gauge correction (O et al., 2017). IMERG-Final is recommended for research applications as being the most accurate but would not be useful for predicting landslides in a timely fashion (Huffman et al., 2020). Since IMERG products use the GPM active and passive microwave data as a standard with little-to-no information from gauges, they are fundamentally different from many other precipitation products available.

For Peer Review

2.3 Precipitation inter-comparison and computation of storm characteristics

For each of the precipitation products, data were extracted for the precipitation grid enclosing the landslide location for the period between May 2015 (the earliest date MRMS data are available) and May 2020 (the latest release of IMERG-Final data at the time of this analysis). Following (Dinku et al., 2008), a minimum threshold of 1 mm/hr was applied to the precipitation data to reduce noise. The data were then split into storm events, where a minimum inter-event time (MIT) criterion of 24 hours as described in Dunkerley (2008) was considered to mark the end of one storm and the beginning of the next.

For each storm, the characteristics of total depth, duration, intensity, and peak intensity were computed and compared. The peak intensity for a storm was the intensity of the single maximum precipitation measurement of the storm. Depth and frequency were chosen since they reflect the most common metrics used in extreme hydrologic events (England et al., 2019). Intensity and duration were included because they are parameters commonly used to study rainfall-triggered landslides (Kirschbaum et al., 2012). Previous studies have suggested that for certain landslides high peak intensity can contribute significantly to triggering a landslide independent of the overall storm depth, duration or intensity (Corominas et al., 2002; Yu et al., 2006). This idea is supported for example by observations that landslides are commonly initialized within hours of the peak intensity (Premchitt et al., 1986). The precipitation rank and z-score among the four products for each landslide event were also computed for the day of the landslide as well as for the full May 2015-May 2020 precipitation record. Rank was chosen as in indicator of the relative magnitude of each product relative to the others, and the z-score as an indicator of the variability of each product relative to the others.

To facilitate comparison of storm characteristics within a single over-arching framework, the return period of the landslide-triggering storms was computed using the NOAA precipitation atlas frequency estimations (US Department of Commerce, 2013). The NOAA atlas provides return periods for discrete precipitation durations, namely 1, 2, 3, 6, 12, 24, 48, 72, 96, and 168 hours. In order to define a consistent return period for each storm, we used the maximum precipitation value for each applicable NOAA atlas duration rather than attempting to expand the storm duration to one of the NOAA atlas durations which might have artificially lowered the return periods. For example, for the 3-hour duration, cumulative 3-hour precipitation totals were calculated for each time step of the storm, and the maximum value chosen. The return period for this maximum value was then retrieved from NOAA atlas. We then selected the maximum return period from among the 10 possible durations noted above for each landslide. For example, if the maximum 3-hour interval during the MIT-defined storm had a 25-year return period while the maximum 48-hour interval during the storm only had a 2-year return period, the return period of the 3-hour return period would be used in preference over the 48-hour return period or any other duration where the maximum return period was

For Peer Review

less than 25 years. This procedure ensured that we used the maximum applicable return period available from the NOAA atlas that occurred during each landslide-triggering storm. Values less than a 2-year return period are not included in the NOAA atlas, such that return period values were only assigned for a subset of landslide-triggering storms that exceeded that threshold. Return period data were also unavailable for Canadian sites.

2.4 Performance of intensity-duration thresholds using different precipitation products

Intensity-Duration thresholds are a category of simple models of landslide occurrence whereby a threshold is defined as a power law of the storm duration

$$I = aD^{-b} \qquad (1)$$

where I is intensity, D is duration, and a and b are fitted parameters to a particular dataset. Intensities above the threshold are used to predict the occurrence of a landslide (Segoni et al., 2014). A range of thresholds have been calculated under different climates and over multiple scales, including globally (Caine, 1980; Kirschbaum et al., 2012; Scheevel et al., 2017). Three thresholds for this study (Caine, 1980; Cannon & Gartner, 2005; Guzzetti et al., 2007) were obtained from a review by Guzzetti et al. (2008) as a way to test the sensitivity of our results to a chosen threshold. Thresholds were only applied to applicable subsets of the data based on climate or other conditions. For example, since the Guzzetti et al. (2007) was defined for "mild, marine west coast climates," only data west of longitude 115W was included in that portion. The Cannon & Gartner (2005) is intended for burned areas, and since the fire history of the locations in this study are unknown it is included only as a comparison point to the other thresholds. For each threshold-product combination, we computed a hit ratio (correctly predicted landslides over the total number of landslides) and a false alarm ratio (incorrectly predicted landslides over the total number of non-landslides)

3 RESULTS

The four precipitation products examined in this study exhibit a great deal of variability in the time period leading up to the landslide event. As an example of the magnitudes and qualitative characteristics of that variability, figure 2 shows the cumulative precipitation in the 30-days before a landslide at five sites. The selected sites showcase multiple ways in which precipitation can differ among the products. For example, while the precipitation in panel (a) matches fairly closely for all products, while in panel (b) precipitation still appears to be correlated but also demonstrates a factor of two spread of precipitation values. In panel (c) the IMERG products diverges substantially from the ground-based products early on in cumulative volume, but the landslide-triggering storm is recorded as nearly twice as large by the satellite-based products, demonstrating that the differences in precipitation measurement can partially cancel out in the right situation. In panel (d) the IMERG-

For Peer Review

Early product reports nearly doubled precipitation values throughout while all three of the remaining products are very similar. Among the events where IMERG-Final recorded a much lower than average value, it was common for the high average to be driven by the IMERG-Early measurements almost exclusively, as shown here in panel (d). Panel (e) shows a likely landslide location error since none of the products register any precipitation close to the time of the event. Such events are not included in most of the analysis of landslide-triggering storms because no such storm could be identified. We note that the differences in precipitation depths accumulated over these 30-day periods are of the same order of magnitude as the *annual* error in depth reported for products of the same category by Sun et al. (2018). This may be because variability among products of different categories, e.g., satellite vs. radar, whereas the figure from Sun et al. (2018) includes only satellite products. Alternatively, when aggregating over a whole year some of the variability among products cancels out, whereas landslide-triggering storms have a greater potential for error by virtue of being relatively brief events.

[Insert Figure 2]

The variability among products is also evident in the distribution of daily precipitation rank among products and z-score within products. The relative magnitude of the different precipitation products on the day of the landslide is shown in Figure 3 in terms of the rank among the four products for each day, and z-score among all non-zero data for a particular product. Both day-of-landslide precipitation and all other non-zero days in the study period are shown for comparison. The ranks of each product do not reveal substantial biases across the entire precipitation record, with the exception of MRMS which has a larger proportion of above median ranks than the other products. On the day of the landslide only, the IMERG products have lower ranks overall, suggested that the satellite-based measurements are less consistent at detecting extreme precipitation. This idea is reinforced by the z-scores of the precipitation among measurements from the product and landslide site. Though the median z-score is similar for all products across the entire record, it is lower for both IMERG products on the day of the landslide. Conversely, some outliers in the IMERG-Early have the highest z-scores among day-of-landslide precipitation even the median and third quartile values are higher, suggesting that the further processing of the IMERG-Final product reduces unusually high precipitation measurements while also increasing low values. For all products, each quartile of the day-of-landslide precipitation is larger than that of the non-landslide-triggering precipitation, though none of the maximum precipitation z-scores appears to have occurred on the day of the landslide.

[Insert Figure 3]

Variability among the precipitation products is also revealed in a comparison of day-of-landslide precipitation with mean values among each of the products. Figure 4 shows the characteristics of the landslide-triggering storms plotted against the ensemble mean of all the products for all the landslide

For Peer Review

sites and separately for the verified locations. There is reasonable agreement among products on the depth and duration of storms, with the exception of outliers below 10mm of total depth—corresponding with a fairly modest storm depth. Among the verified locations, there are fewer low-depth or duration values that are either outliers or near to the mean, suggesting that low measurements may reflect limitations in the GLC location accuracy for sites with only approximate locations.

The two satellite products have a pattern of distinctive readings relative to ground-based products in that they contain both the highest and lowest values of several storm characteristics. The IMERG products generally report higher peak hourly intensities for the storms with the highest mean peak intensity, which is likely at least partially due to the shorter 30-minute time step. The higher peak intensities are also reflected to some degree in longer return periods, which are based on hourly durations or longer for comparison with the NOAA Atlas. An examination of the relationship between return period and peak intensity showed a clear relationship, with the return period data reflecting a subset of the highest intensity storms due to the 2-year return period cut-off. The relationship between peak intensity and return period is not surprising given that the return period values were calculated by searching for the most intense period of each NOAA atlas duration. However, IMERG-Final has lower return periods over all when all landslide locations are included despite reporting high peak intensities. This anomaly disappears in an examination of verified locations alone. An examination of the 30-day precipitation record prior to the landslide for sites where the IMERG-Final return period was much lower than the average revealed that in most of those cases the higher mean was driven primarily by anomalously high IMERG-Early values not reflected in any of the other datasets. MRMS and NLDAS-2 have lower return periods that the IMERG products even among the verified locations only, suggesting that these products may not consistently detect the highest return period precipitation events.

[Insert Figure 4]

The precipitation products are examined in the context of landslide triggering thresholds in Figure 5, with the performance summarized in Table 2. Interestingly, the choice of intensity-duration threshold does not appear to make a large difference in performance because the threshold curves are more similar than the variation in the precipitation data across sites and among products. The MRMS or NLDAS-2 products tend to perform better than either IMERG product, with hit ratios between 0.84-0.88 and 0.65-0.76 rather than 0.59-0.61 and 0.59-0.73 among the verified landslide locations, respectively. All products perform comparably or better when using only the verified landslide locations relative to the approximate locations

Figure 5 shows a concentration of long-duration, low-intensity storms that are in the vicinity of 24-hour duration for all products. These storms may be an artifact of the MIT storm identification algorithm. Since the landslides did not have times specified, the entire day of the landslide was always

For Peer Review

included unless there was no rain until the end of the day, and this may have extended some storms past when the landslide occurred. This would have the effect of computing lower total intensity values for storms that lasted only through the time of the landslide but persisted only at a much lower intensity thereafter. Many of the storms that did trigger landslides but were not correctly identified by the intensity-duration threshold fall into this group of approximately 24-hour low-intensity storms. Adjustments to storm delineation through a different algorithm or a higher minimum threshold may increase performance, especially for the IMERG products which showed the most low-intensity landslide triggering storms.

[Insert Figure 5]

[Insert Table 2]

4 DISCUSSION

Among the precipitation products chosen for this study, the two IMERG products identify both higher peak intensities and longer return periods relative to the other products. Interestingly, they also detect more anomalously low precipitation values. Low-intensity precipitation in all products was associated with long duration storm events (see figure 5)), which may occur because of "drizzle," i.e., low-intensity precipitation slightly above the 1 mm threshold that extended the MIT-computed duration of the storm and thereby reduced its overall intensity. As a result, the IMERG products were particularly vulnerable to the identification of long-duration low-intensity storms as a result of the MIT method used in this study to separate storms. Those long-duration low-intensity storms had the effect of lowering the hit ratio. Because the IMERG products were able to identify higher intensity precipitation than the other products, it is possible that they would in fact perform better for identifying landslides if the low-intensity storm problem were mitigated.

All precipitation products performed reasonably well at identifying landslides using the published intensity-duration thresholds particularly considering that these thresholds were developed on training data from different datasets spanning large regions. However, they did not perform as well at excluding false alarms, most likely because of factors beyond intensity and duration that can influence landslide occurrence such as topography, soil type, recent wildfire or disturbance or land development. Some of the high-intensity precipitation that did not trigger any recorded landslides could be more reflective of adjacent areas that are not as susceptible to landslides. Conversely a landslide at a highly susceptible location, such as an area with high slopes that had recently been burned by wildfire could be triggered by less intense rain, potentially resulting as a miss on an intensity-duration curve. Even the 1.1 km resolution of the MRMS data could contain substantial variation in landslide susceptibility within an individual grid cell. The poorest performing products were the IMERG products because despite their detecting more high-intensity events they also

For Peer Review

detected many low-intensity long-duration events, causing the intensity-duration threshold to miss landslides.

Both Rossi et al. (2017) and Brunetti et al. (2018) also found satellite products did not perform as well as gauge data relative to intensity-duration thresholds, as a result of underestimating precipitation. However, those studies found that adjusting the threshold accounted for precipitation bias, suggesting that the intensities were lower in a more or less uniform pattern across different durations. By contrast, in this study we found that the low-intensity values were more often clustered around a relatively small duration band, which would be more challenging to bias-correct. Though the intensity-duration thresholds still show promise for diagnosing landslides using satellite-based data, the adjustments to improve performance may prove more complex for the IMERG products across the broader spatial domain of the continental US and Canada.

MRMS and NLDAS-2 are relatively low latency products. In the case of IMERG-Early the short latency seemed to come at a cost of an exaggeration of the weaknesses and strengths of IMERG in identifying landslides. In particular, IMERG-Early had the greatest prevalence of low storm intensities, and so it ultimately performed the worst at landslide identification. Without changes to the precipitation processing, the low latency does indeed appear to be a liability in this case.

Precipitation measurements at verified landslide sites tended to be of higher magnitude than those at other sites with approximate locations for all products. The intensity-duration thresholds subsequently performed better at verified locations across all precipitation products. Though this difference remains unexplained, one possibility is that some of the approximate landslide locations were too far away from the true landslide location for the precipitation measurements to be representative. Alternatively, there may have been other factors such as vegetation cover that made it more difficult to locate landslides on satellite imagery and also lowered the precipitation threshold that would trigger a landslide. Since work on this study began, a compilation of U.S. landslides has been released by the USGS (Mirus et al., 2020) which would also be a suitable source of landslide locations with perhaps greater location precision that could help resolve this question in future work along the same lines.

5 CONCLUSION

The precipitation products chosen for this study represent diverse measurement techniques that often recorded large differences in precipitation leading up to the landslide events evaluated here. As a result, each precipitation product differed in overall performance in predicting landslides using intensity-duration thresholds. Overall, the choice of intensity-duration threshold was not as consequential as the choice of precipitation product in identifying landslides. Performance from products that rely on ground-based sensors showed a more consistent landslide signal despite generally recording lower peak intensities and return periods.

For Peer Review

Though it was hypothesized that peak intensity would be an important factor in identifying landslides, the results suggest instead that removal of noise on the low end, i.e., drizzle, may be more important. A particular challenge was the presence of low-intensity, long-duration storms preceding landslide events, most prevalent in the IMERG products. A more expansive evaluation of processing techniques for separating storms may potentially mitigate these issue, although each technique will produce artifacts in the comparisons. Another potential avenue for addressing this problem is to combine multiple dataset, since the low-intensity long-duration storms did not appear in all datasets to the same degree.

Another limitation to the study landslide-triggering storms is the general lack of both exact landslide locations and specific time of day of the landslide events. The location limitation was reflected in better performance for verified landslide locations as compared to approximate locations, which implies that some of the approximate locations were incorrect to such an extent that the precipitation measurements were misaligned. This problem could be addressed by more extensive manual searches such as the one used in this study that identified the 80 verified landslide locations, or perhaps in the future by machine learning methods. Since work on this study began, a compilation of U.S. landslides has been released by the USGS (Mirus et al., 2020) which would also be a suitable source of landslide locations with perhaps greater location precision that could be used in future work along the same lines.

Using the methods tested in this study, those practitioners attempting to use intensity-duration thresholds as operation landslide models would do well to select a product like MRMS that has extremely low latency and performs well at identifying landslides. None of the products were particularly good at filtering out false alarms of landslides. Therefore, an additional recommendation would be for practitioners to consider more than one precipitation product, i.e., multiple precipitation estimates simultaneously, as a way to confirm stronger precipitation signals and to minimize the influence of noise.

DATA AVAILABILITY

These data were derived from the following resources available in the public domain: The NASA Global Landslide Catalog was downloaded from NASA's Open Data Portal (https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog/h9d8-neg4); Both IMERG products were retrieved from the Global Precipitation Measurement data portal (https://gpm.nasa.gov/data/directory); NLDAS-2 data is available on NASA's EarthData site (https://disc.gsfc.nasa.gov/datasets/NLDAS_FORA0125_H_002/summary?keywords=NLDAS/); and MRMS data was retrieved from a public archive at Iowa State University (https://mtarchive.geol.iastate.edu/). The verified locations of GLC landslides used in this study are available from the corresponding author upon reasonable request.

For Peer Review

REFERENCES

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., & Nelkin, E. (2003). The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979Present). *Journal of Hydrometeorology*, *4*(6), 1147–1167. https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2

Adler, R. F., Kidd, C., Petty, G., Morissey, M., & Goodman, H. M. (2001). Intercomparison of Global Precipitation Products: The Third Precipitation Intercomparison Project (PIP-3). *Bulletin of the American Meteorological Society*, 20.

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., & Amitai, E. (2011). Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *Journal of Geophysical Research: Atmospheres*, *116*(D2). https://doi.org/10.1029/2010JD014741

Ahmadalipour, A., & Moradkhani, H. (2017). Analyzing the uncertainty of ensemble-based gridded observations in land surface simulations and drought assessment. *Journal of Hydrology*, *555*, 557–568. https://doi.org/10.1016/j.jhydrol.2017.10.059

Amitai, E., Petersen, W., Llort, X., & Vasiloff, S. (2012). Multiplatform Comparisons of Rain Intensity for Extreme Precipitation Events. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(3), 675–686. https://doi.org/10.1109/TGRS.2011.2162737

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., & Prat, O. P. (2015). PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies. *Bulletin of the American Meteorological Society*, *96*(1), 69–83. https://doi.org/10.1175/BAMS-D-13-00068.1

Bao, J., Sherwood, S. C., Alexander, L. V., & Evans, J. P. (2017). Future increases in extreme precipitation exceed observed scaling rates. *Nature Climate Change*, *7*(2), 128–132. https://doi.org/10.1038/nclimate3201

Beck, H. E., van Dijk, A. I., Levizzani, V., Schellekens, J., Miralles, D., Martens, B., & de Roo, A. (2017). MSWEP: 3-hourly 0.25 global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *HYDROLOGY AND EARTH SYSTEM SCIENCES*, *21*(1), 589–615. https://doi.org/10.5194/hess-21-589-2017

Bousquet, O., & Smull, B. F. (2003). Observations and impacts of upstream blocking during a widespread orographic precipitation event. *Quarterly Journal of the Royal Meteorological Society*, *129*(588), 391–409. https://doi.org/10.1256/qj.02.49

For Peer Review

Caine, N. (1980). The Rainfall Intensity - Duration Control of Shallow Landslides and Debris Flows. *Geografiska Annaler: Series A, Physical Geography*, *62*(1-2), 23–27. https://doi.org/10.1080/04353676.1980.11879996

Cannon, S. H., & Gartner, J. E. (2005). Wildfire-related debris flow from a hazards perspective. In *Debris-flow Hazards and Related Phenomena* (p. 23).

Cannon, S. H., Gartner, J. E., Wilson, R. C., Bowers, J. C., & Laber, J. L. (2008). Storm rainfall conditions for floods and debris flows from recently burned areas in southwestern Colorado and southern California. *Geomorphology*, *96*(3), 250–269. https://doi.org/10.1016/j.geomorph.2007.03.019

Chandrasekar, V., Hou, A., Smith, E., Bringi, V. N., Rutledge, S. A., Gorgucci, E., Petersen, W. A., & Jackson, G. S. (2008). Potential Role of Dual-Polarization Radar in the Validation of Satellite Precipitation Measurements: Rationale and Opportunities. *Bulletin of the American Meteorological Society*, *89*(8), 1127–1146. https://doi.org/10.1175/2008BAMS2177.1

Chowdhury, R., & Flentje, P. (2002). Uncertainties in rainfall-induced landslide hazard. *Quarterly Journal of Engineering Geology and Hydrogeology*, *35*(1), 61–69. https://doi.org/10.1144/qjegh.35.1.61

Clarizia, M., Gullà, G., & Sorbino, G. (1996). Sui meccanismi di innesco dei soil slip. *International Conference Prevention of Hydrogeological Hazards: The Role of Scientific Research*, *1*, 585–597.

Corominas, J., Moya, J., & Hürlimann, M. (2002). Landslide Rainfall Triggers in the Spanish Eastern Pyrenees. *Mediterranean Storms*, 4.

Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S. J., & Ropelewski, C. F. (2008). Validation of high-resolution satellite rainfall products over complex terrain. *International Journal of Remote Sensing*, *29*(14), 4097–4110. https://doi.org/10.1080/01431160701772526

Duchon, C. E., & Biddle, C. J. (2010). Undercatch of tipping-bucket gauges in high rain rate events. *Advances in Geosciences*, *25*, 11–15. https://doi.org/10.5194/adgeo-25-11-2010

Duchon, C., Fiebrich, C., & Grimsley, D. (2014). Using High-Speed Photography to Study Undercatch in Tipping-Bucket Rain Gauges. *Journal of Atmospheric and Oceanic Technology*, *31*(6), 1330–1336. https://doi.org/10.1175/JTECH-D-13-00169.1

Dunkerley, D. (2008). Identifying individual rain events from pluviograph records: A review with analysis of data from an Australian dryland site. *Hydrological Processes*, *22*(26), 5024–5036. https://doi.org/10.1002/hyp.7122

For Peer Review

Ebert, E. E. (2007). Methods for Verifying Satellite Precipitation Estimates. In V. Levizzani, P. Bauer, & F. J. Turk (Eds.), *Measuring Precipitation From Space* (pp. 345–356). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5835-6_27

England, J. F., Jr., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas Jr., W. O., Veilleux, A. G., Kiang, J. E., & Mason, R. R., Jr. (2019). Guidelines for determining flood flow frequency Bulletin 17C. In *Guidelines for determining flood flow frequency Bulletin 17C* (USGS Numbered Series No. 4-B5; Techniques and Methods, p. 168). U.S. Geological Survey. https://doi.org/10.3133/tm4B5

Fornasiero, A., Amorati, R., Alberoni, P. P., Ferraris, L., & Taramasso, A. C. (2004). Impact of combined beam blocking and anomalous propagation correction algorithms on radar data quality. *Proceedings of ERAD*, 216–222.

Froude, M. J., & Petley, D. N. (2018). Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, *18*(8), 2161–2181. https://doi.org/10.5194/nhess-18-2161-2018

Galanti, Y., Barsanti, M., Cevasco, A., D'Amato Avanzi, G., & Giannecchini, R. (2018). Comparison of statistical methods and multi-time validation for the determination of the shallow landslide rainfall thresholds. *Landslides*, *15*(5), 937–952. https://doi.org/10.1007/s10346-017-0919-3

Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A., & Rasmussen, R. M. (2014). An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, *50*(9), 7167–7186. https://doi.org/10.1002/2014WR015559

Guzzetti, F., Peruccacci, S., Rossi, M., & Stark, C. P. (2008). The rainfall intensity duration control of shallow landslides and debris flows: An update. *Landslides*, *5*(1), 3–17. https://doi.org/10.1007/s10346-007-0112-1

Guzzetti, F., Peruccacci, S., Rossi, M., & Stark, C. P. (2007). Rainfall thresholds for the initiation of landslides in central and southern Europe. *Meteorology and Atmospheric Physics*, *98*(3-4), 239–267. https://doi.org/10.1007/s00703-007-0262-7

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Highland, L., & Bobrowsky, P. (2008). *The Landslide Handbook: A Guide to Understanding Landslides* (Circular No. 1325; Circular, p. 129). United States Geological Survey.

For Peer Review

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., & Iguchi, T. (2014). The Global Precipitation Measurement Mission. *Bulletin of the American Meteorological Society*, *95*(5), 701–722. https://doi.org/10.1175/BAMS-D-13-00164.1

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K.-L., Joyce, R. J., Kidd, C., Nelkin, E. J., Sorooshian, S., Stocker, E. F., Tan, J., Wolff, D. B., & Xie, P. (2020). Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG). In V. Levizzani, C. Kidd, D. B. Kirschbaum, C. D. Kummerow, K. Nakamura, & F. J. Turk (Eds.), *Satellite Precipitation Measurement: Volume 1* (pp. 343–353). Springer International Publishing. https://doi.org/10.1007/978-3-030-24568-9_19

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., & Stocker, E. F. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *Journal of Hydrometeorology*, *8*(1), 38–55. https://doi.org/10.1175/JHM560.1

Janssen, E., Wuebbles, D. J., Kunkel, K. E., Olsen, S. C., & Goodman, A. (2014). Observational- and model-based trends and projections of extreme precipitation over the contiguous United States. *Earth's Future*, *2*(2), 99–113. https://doi.org/10.1002/2013EF000185

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., & Kirschbaum, D. B. (2017). So, how much of the Earth's surface is covered by rain gauges? *Bulletin of the American Meteorological Society*, *98*(1), 69–78. https://doi.org/10.1175/BAMS-D-14-00283.1

Kidd, C., Takayabu, Y. N., Skofronick-Jackson, G. M., Huffman, G. J., Braun, S. A., Kubota, T., & Turk, F. J. (2020). The Global Precipitation Measurement (GPM) Mission. In V. Levizzani, C. Kidd, D. B. Kirschbaum, C. D. Kummerow, K. Nakamura, & F. J. Turk (Eds.), *Satellite Precipitation Measurement* (Vol. 67, pp. 3–23). Springer International Publishing. https://doi.org/10.1007/978-3-030-24568-9_1

Kirschbaum, D. B., Adler, R., Hong, Y., Hill, S., & Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: Method, results, and limitations. *Natural Hazards*, *52*(3), 561–575. https://doi.org/10.1007/s11069-009-9401-4

Kirschbaum, D. B., Adler, R., Hong, Y., Kumar, S., Peters-Lidard, C., & Lerner-Lam, A. (2012). Advances in landslide nowcasting: Evaluation of a global and regional modeling approach. *Environmental Earth Sciences*, *66*(6), 1683–1696. https://doi.org/10.1007/s12665-011-0990-3

Kirschbaum, D., & Stanley, T. (2018). Satellite-Based Assessment of Rainfall-Triggered Landslide Hazard for Situational Awareness. *Earth's Future*, *6*(3), 505–523. https://doi.org/10.1002/2017EF000715

For Peer Review

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., & Simpson, J. (1998). The Tropical Rainfall Measuring Mission (TRMM) Sensor Package. *Journal of Atmospheric and Oceanic Technology*, *15*(3), 809–817. https://doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2

Leonarduzzi, E., Molnar, P., & McArdell, B. W. (2017). Predictive performance of rainfall thresholds for shallow landslides in Switzerland from gridded daily data. *Water Resources Research*, *53*(8), 6612–6625. https://doi.org/10.1002/2017WR021044

Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., Cayan, D. R., & Brekke, L. (2015). A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950. *Scientific Data*, *2*(1), 150042. https://doi.org/10.1038/sdata.2015.42

Lockhoff, M., Zolina, O., Simmer, C., & Schulz, J. (2014). Evaluation of Satellite-Retrieved Extreme Precipitation over Europe using Gauge Observations. *Journal of Climate*, *27*(2), 607–623. https://doi.org/10.1175/JCLI-D-13-00194.1

Long, D., Longuevergne, L., & Scanlon, B. R. (2014). Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research*, *50*(2), 1131–1151. https://doi.org/10.1002/2013WR014581

Lundquist, J. D., Hughes, M., Henn, B., Gutmann, E. D., Livneh, B., Dozier, J., & Neiman, P. (2015). High-Elevation Precipitation Patterns: Using Snow Measurements to Assess Daily Gridded Datasets across the Sierra Nevada, California. *Journal of Hydrometeorology*, *16*(4), 1773–1792. https://doi.org/10.1175/JHM-D-15-0019.1

Manzanas, R., Amekudzi, L. K., Preko, K., Herrera, S., & Gutiérrez, J. M. (2014). Precipitation variability and trends in Ghana: An intercomparison of observational and reanalysis products. *Climatic Change*, *124*(4), 805–819. https://doi.org/10.1007/s10584-014-1100-9

Mirus, B. B., Jones, E. S., Baum, R. L., Godt, J. W., Slaughter, S., Crawford, M. M., Lancaster, J., Stanley, T., Kirschbaum, D. B., Burns, W. J., Schmitt, R. G., Lindsey, K. O., & McCoy, K. M. (2020). Landslides across the USA: Occurrence, susceptibility, and data limitations. *Landslides*, *17*(10), 2271–2285. https://doi.org/10.1007/s10346-020-01424-4

Nikahd, A., Hashim, M., & Nazemosadat, M. J. (2016). A Review of Uncertainty Sources on Weather Ground-Based Radar for Rainfall Estimation. In *Applied Mechanics and Materials* (Vol. 818, pp. 254–271). https://www.scientific.net/AMM.818.254; Trans Tech Publications Ltd. https://doi.org/10.4028/www.scientific.net/AMM.818.254

North America Elevation 1-Kilometer Resolution (Third). (2007). [Map]. Commission for Environmental Cooperation.

For Peer Review

O, S., Foelsche, U., Kirchengast, G., Fuchsberger, J., Tan, J., & Petersen, W. A. (2017). Evaluation of GPM IMERG Early, Late, and Final rainfall estimates using WegenerNet gauge data in southeastern Austria. *Hydrology and Earth System Sciences*, *21*(12), 6559–6572. https://doi.org/10.5194/hess-21-6559-2017

Pendergrass, A. G., & Knutti, R. (2018). The Uneven Nature of Daily Precipitation and Its Change. *Geophysical Research Letters*, *45*(21), 11, 980–911, 988. https://doi.org/10.1029/2018GL080298

Pollock, M. D., O'Donnell, G., Quinn, P., Dutton, M., Black, A., Wilkinson, M. E., Colli, M., Stagnaro, M., Lanza, L. G., Lewis, E., Kilsby, C. G., & O'Connell, P. E. (2018). Quantifying and Mitigating Wind-Induced Undercatch in Rainfall Measurements. *Water Resources Research*, *54*(6), 3863–3875. https://doi.org/10.1029/2017WR022421

Premchitt, J., Brand, E. W., & Phillipson, H. B. (1986). Landslides caused by rapid groundwater changes. *Geological Society, London, Engineering Geology Special Publications*, *3*(1), 87–94. https://doi.org/10.1144/GSL.ENG.1986.002.01.09

Rossi, M., Kirschbaum, D., Valigi, D., Mondini, A. C., & Guzzetti, F. (2017). Comparison of Satellite Rainfall Estimates and Rain Gauge Measurements in Italy, and Impact on Landslide Modeling. *Climate*, *5*(4), 90. https://doi.org/10.3390/cli5040090

Scheevel, C. R., Baum, R. L., Mirus, B. B., & Smith, J. B. (2017). *Precipitation Thresholds for Landslide Occurrence Near Seattle, Mukilteo, and Everett, Washington* (Open-File Report No. 2017-1039; Open-File Report). U.S. Department of the Interior; U.S. Geological Survey.

Segoni, S., Rossi, G., Rosi, A., & Catani, F. (2014). Landslides triggered by rainfall: A semi-automated procedure to define consistent intensity duration thresholds. *Computers & Geosciences*, *63*, 123–131. https://doi.org/10.1016/j.cageo.2013.10.009

Skofronick-Jackson, G., Petersen, W. A., Berg, W., Kidd, C., Stocker, E. F., Kirschbaum, D. B., Kakar, R., Braun, S. A., Huffman, G. J., Iguchi, T., Kirstetter, P. E., Kummerow, C., Meneghini, R., Oki, R., Olson, W. S., Takayabu, Y. N., Furukawa, K., & Wilheit, T. (2017). The Global Precipitation Measurement (GPM) Mission for Science and Society. *Bulletin of the American Meteorological Society*, *98*(8), 1679–1695. https://doi.org/10.1175/BAMS-D-15-00306.1

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K.-L. (2018). A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics*, *56*(1), 79–107. https://doi.org/10.1002/2017RG000574

Sunyer, M. A., Hundecha, Y., Lawrence, D., Madsen, H., Willems, P., Martinkova, M., Vormoor, K., Bürger, G., Hanel, M., Kriaučiūnienė, J., Loukas, A., Osuch, M., & Yücel, I. (2015). Inter-comparison

For Peer Review

of statistical downscaling methods for projection of extreme precipitation in Europe. *Hydrology and Earth System Sciences*, *19*(4), 1827–1847. https://doi.org/10.5194/hess-19-1827-2015

Tapiador, F. J., Turk, F. J., Petersen, W., Hou, A. Y., García-Ortega, E., Machado, L. A. T., Angelis, C. F., Salio, P., Kidd, C., Huffman, G. J., & de Castro, M. (2012). Global precipitation measurement: Methods, datasets and applications. *Atmospheric Research*, *104-105*, 70–97. https://doi.org/10.1016/j.atmosres.2011.10.021

Tryhorn, L., & DeGaetano, A. (2011). A comparison of techniques for downscaling extreme precipitation over the Northeastern United States. *International Journal of Climatology*, *31*(13), 1975–1989. https://doi.org/10.1002/joc.2208

US Department of Commerce, N. (2013). NOAA Atlas 2 Precipitation Frequency Estimates in GIS Compatible Formats. In *NOAA Atlas 2 Precipitation Frequency Estimates in GIS Compatible Formats*. https://www.nws.noaa.gov/ohd/hdsc/noaaatlas2.htm; US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, M. J., Williams, C. N., Fenimore, C., Gleason, K., & Arndt, D. (2014). Improved Historical Temperature and Precipitation Time Series for U.S. Climate Divisions. *Journal of Applied Meteorology and Climatology*, *53*(5), 1232–1251. https://doi.org/10.1175/JAMC-D-13-0248.1

Wang, G., Kirchhoff, C. J., Seth, A., Abatzoglou, J. T., Livneh, B., Pierce, D. W., Fomenko, L., & Ding, T. (2020). Projected Changes of Precipitation Characteristics Depend on Downscaling Method and Training Data: MACA versus LOCA Using the U.S. Northeast as an Example. *JOURNAL OF HYDROMETEOROLOGY*, *21*, 20.

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Brewer, M., Mocko, D., Kumar, S. V., Wei, H., Meng, J., & Luo, L. (2016). Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research: Atmospheres*, *121*(6), 2750–2779. https://doi.org/10.1002/2015JD023733

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., & Mocko, D. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, *117*(D3). https://doi.org/10.1029/2011JD016048

For Peer Review

Yu, F.-C., Chen, T.-C., Lin, M.-L., Chen, C.-Y., & Yu, W.-H. (2006). Landslides and Rainfall Characteristics Analysis in Taipei City during the Typhoon Nari Event. *Natural Hazards*, *37*(1-2), 153–167. https://doi.org/10.1007/s11069-005-4661-0

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cocks, S., Martinaitis, S., Arthur, A., Cooper, K., Brogden, J., & Kitzmiller, D. (2015). Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1

For Peer Review

TABLES

Table 1 | The four precipitation products included in the comparison, representing gauge-, radar-, and satellite-based measurements.

Precipitation product	Description	Spatial Resolution	Temporal resolution	Typical Latency
Integrated Multi-satellitE Retrievals for Global precipitation measurement early run (IMERG-Early; Hou et al., 2014)	Global network of satellites unified by measurements from a single reference radar/radiometer satellite.	0.1° (~10 km)	30 minutes	4 hours
Integrated Multi-satellitE Retrievals for Global precipitation measurement (IMERG-Final; Hou et al., 2014)	In addition to the satellite data included in IMERG-Early, IMERG-Final includes late-arriving microwave overpasses, monthly gauge-based adjustments, and an algorithm that interpolates forward as well as backward in time.	0.1° (~10 km)	30 minutes	3.5 months
Multi-Radar Multi-Sensor (MRMS; Zhang et al., 2015)	Integrates data from radars, satellites, precipitation gages, and other sensors to provide near real-time decision support.	0.01° (~1.1 km)	1 hour	< 5 minutes
North American Land Data Assimilation System version 2 (NLDAS-2) meteorological (Xia et al., 2012)	Disaggregation of Climate Prediction Center daily precipitation using bias-corrected radar	0.125° (~ 12 km)	1 hour	4 days

For Peer Review

744 **Table 2 |** Hit ratio and false alarm ratio for each product and the Guzzetti et al. (2008), Clarizia et al.
745 (1996), and Cannon et al. (2008) intensity-duration thresholds.

Product	Include sites	**Hit ratio**			**False alarm ratio**		
Intensity-duration threshold:		Guzzetti et al. (2008)	Clarizia et al. (1996)	Cannon et al. (2008)	Guzzetti et al. (2008)	Clarizia et al. (1996)	Cannon et al. (2008)
IMERG-Early	All (n=146)	0.56	0.54	0.47	0.17	0.16	0.11
	Verified (n=59)	0.61	0.59	0.54	0.19	0.18	0.12
IMERG-Final	All (n=145)	0.61	0.61	0.53	0.20	0.20	0.13
	Verified (n=56)	0.73	0.75	0.59	0.23	0.23	0.16
NLDAS-2	All (n=131)	0.73	0.74	0.68	0.22	0.24	0.19
	Verified (n=55)	0.76	0.76	0.65	0.22	0.24	0.19
MRMS	All (n=134)	0.82	0.83	0.78	0.24	0.26	0.21
	Verified (n=56)	0.88	0.86	0.84	0.26	0.28	0.22

746

747

For Peer Review

FIGURE LEGENDS

Figure 1 | Map of all landslide sites considered in this analysis: 228 landslide sites colored by whether the location was approximate (n=148) or verified using aerial satellite imagery to identify a visible scarp (n=80); Source of landslide locations was the GLC (Kirschbaum et al., 2010), source of the DEM data used for the base map (*North America Elevation 1-Kilometer Resolution*, 2007). Elevations above 3000 m are indicated as the highest value on the color scale.

Figure 2 | Exposition into the types of precipitation differences leading up to landslide events: Cumulative precipitation measurements at select landslide sites for the 30 days before the event. The precipitation is variable across the different products, and the selected sites each demonstrate diverse types of variability. Panel (a) shows similar measurements among all products throughout the 30 days. In panel (b), all products are well correlated, but the accumulated depths greatly differ. In panel (c) both IMERG products reports less precipitation until the landslide-triggering storm when they reverse and report more precipitation. In panel (d) IMERG-Early reports much more precipitation than the other products. Finally, in panel (e) no landslide-triggering precipitation was detected by any product, suggesting a location error in the landslide record.

Figure 3 | Relative magnitude of precipitation products on the day of the landslide: Rank among all products for each day, and z-score of daily precipitation as measured by each product for each of 228 events. Panels (a) and (c) show the entire precipitation record while panels (b) and (d) show only the day-of-landslide precipitation for comparison. Z-score are plotted on a pseudo-log scale, a combination of a linear scale near zero and a log scale for higher values.

Figure 4 | Storm characteristics vs. the ensemble mean: Depth, duration, intensity, peak intensity, and return period for each of the landslide-triggering storms as measured by four precipitation products. Least-squares regression lines with 95% confidence intervals are also shown. Panels (a)-(e) show all 228 sites while the bottom panel only shows the 80 verified locations.

Figure 5 | Comparison of landslide-triggering precipitation relative to intensity-duration thresholds: Each storm in the precipitation record and established global or climactic intensity-duration thresholds. Landslide-triggering storms are shaded with darker blue. The panels (a)-(d)

For Peer Review

contains precipitation data for all sites while in panels (e)-(h) only verified sites are included. Points above each threshold are predicted by the threshold to be landslides, and so a larger proportion of landslides above the threshold indicates better performance.

For Peer Review

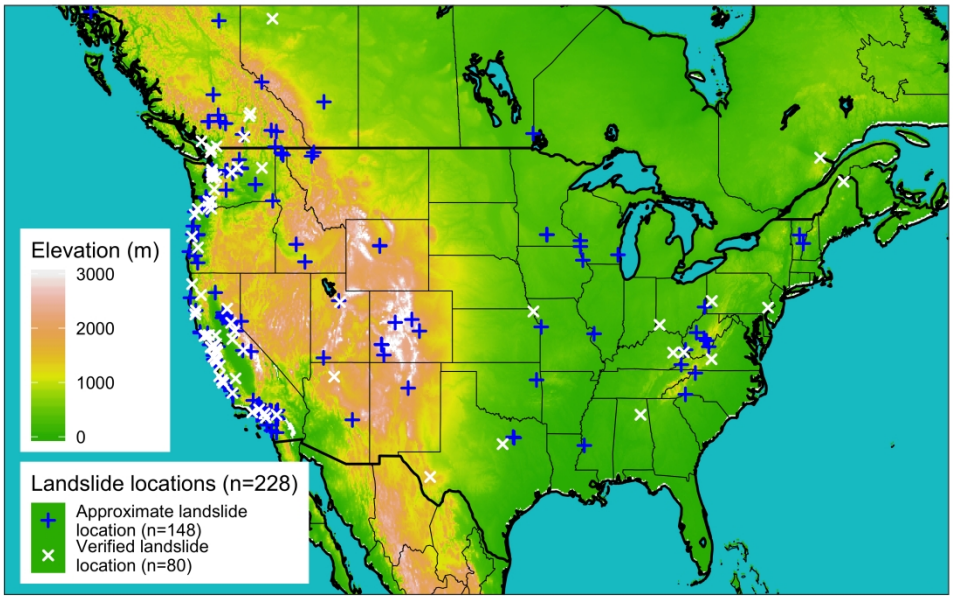


Figure 1 | Map of all landslide sites considered in this analysis: 228 landslide sites colored by whether the location was approximate (n=148) or verified using aerial satellite imagery to identify a visible scarp (n=80); Source of landslide locations was the GLC (Kirschbaum et al., 2010), source of the DEM data used for the base map (North America Elevation 1-Kilometer Resolution, 2007). Elevations above 3000 m are indicated as the highest value on the color scale.

1481x952mm (72 x 72 DPI)

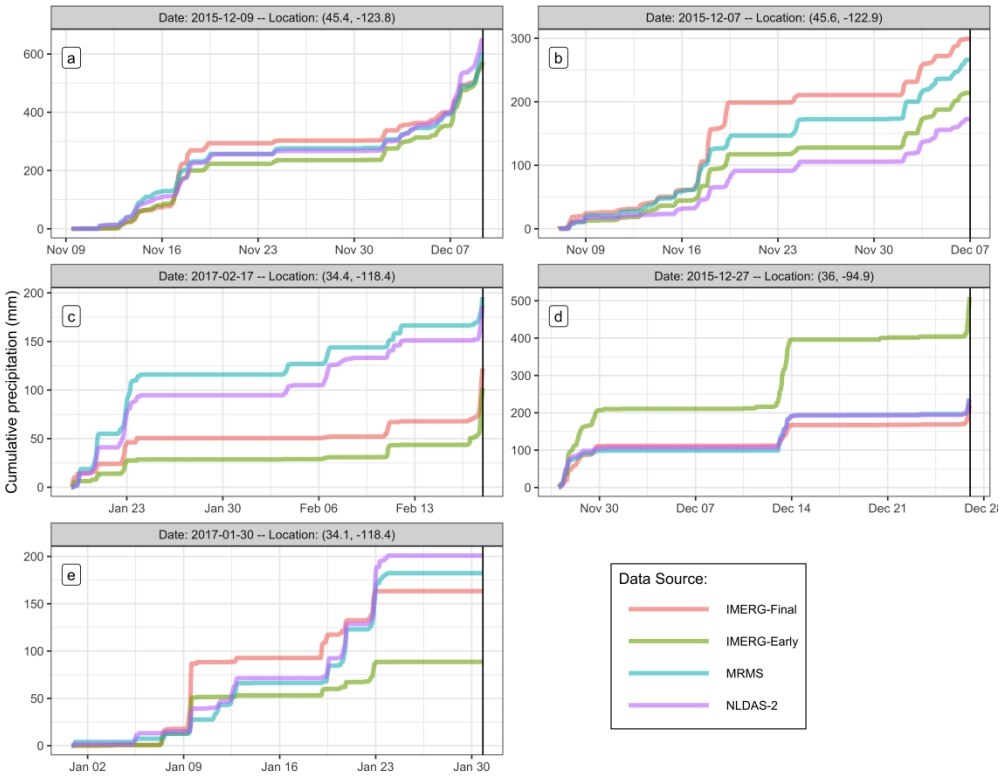


Figure 2 | Exposition into the types of precipitation differences leading up to landslide events: Cumulative precipitation measurements at select landslide sites for the 30 days before the event. The precipitation is variable across the different products, and the selected sites each demonstrate diverse types of variability. Panel (a) shows similar measurements among all products throughout the 30 days. In panel (b), all products are well correlated, but the accumulated depths greatly differ. In panel (c) both IMERG products reports less precipitation until the landslide-triggering storm when they reverse and report more precipitation. In panel (d) IMERG-Early reports much more precipitation than the other products. Finally, in panel (e) no landslide-triggering precipitation was detected by any product, suggesting a location error in the landslide record.

2116x1693mm (72 x 72 DPI)

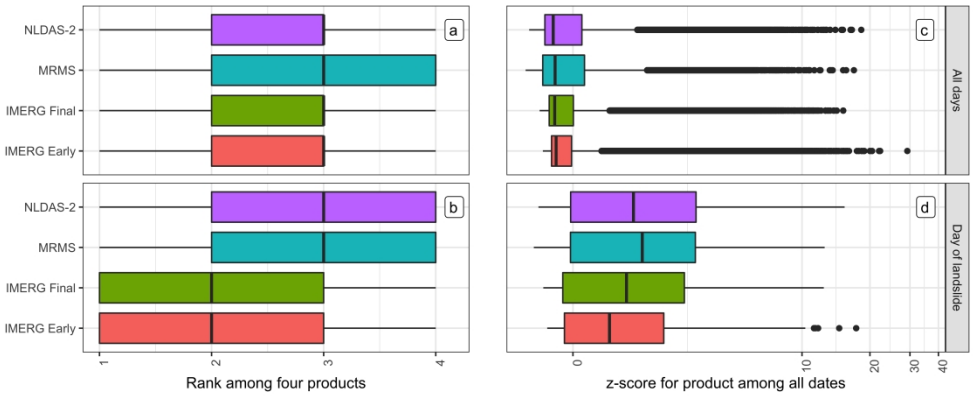


Figure 3 | Relative magnitude of precipitation products on the day of the landslide: Rank among all products for each day, and z-score of daily precipitation as measured by each product for each of 228 events. Panels (a) and (c) show the entire precipitation record while panels (b) and (d) show only the day-of-landslide precipitation for comparison. Z-score are plotted on a pseudo-log scale, a combination of a linear scale near zero and a log scale for higher values.

2116x846mm (72 x 72 DPI)

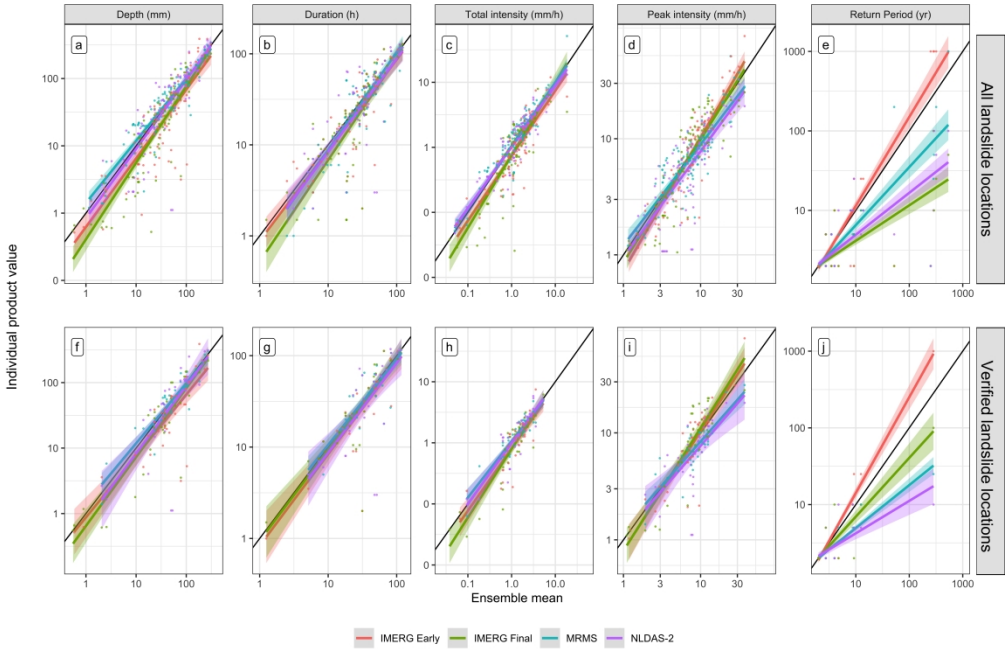


Figure 4 | Storm characteristics vs. the ensemble mean: Depth, duration, intensity, peak intensity, and return period for each of the landslide-triggering storms as measured by four precipitation products. Least-squares regression lines with 95% confidence intervals are also shown. Panels (a)-(e) show all 228 sites while the bottom panel only shows the 80 verified locations.

2540x1693mm (72 x 72 DPI)

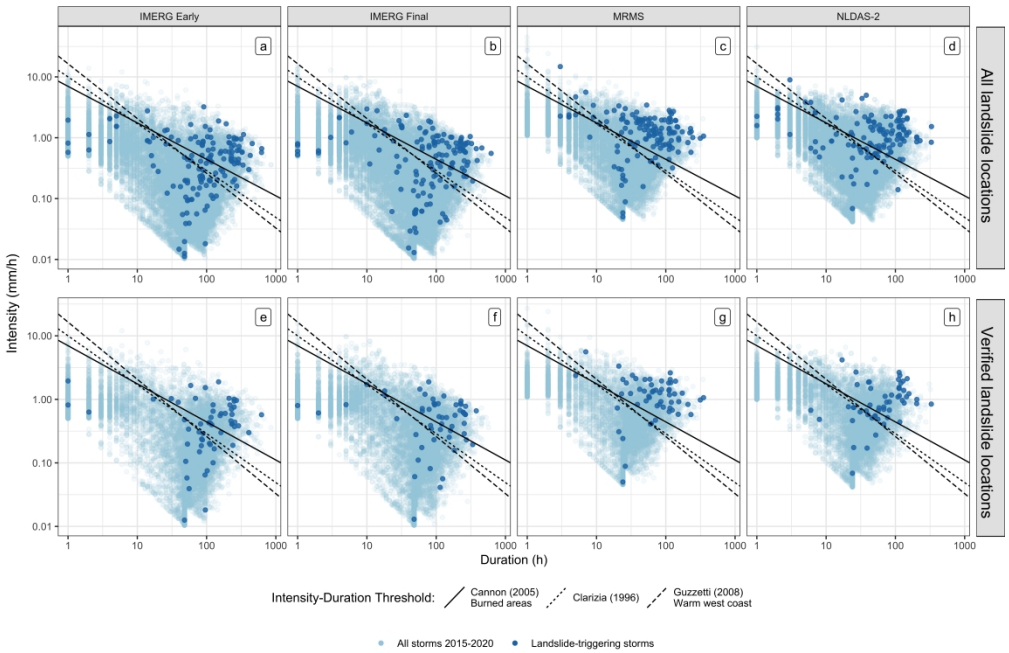
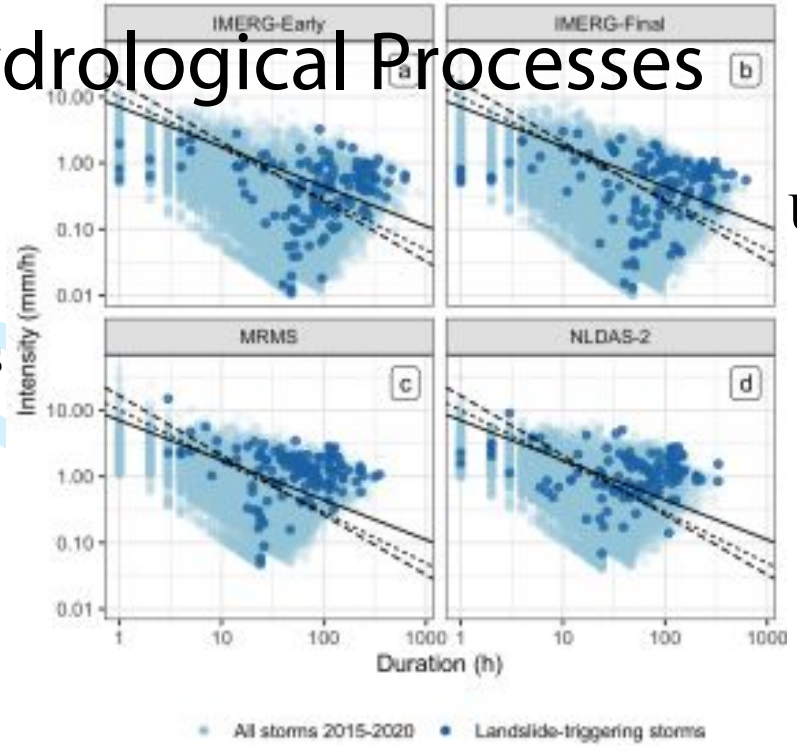


Figure 5 | Comparison of landslide-triggering precipitation relative to intensity-duration thresholds: Each storm in the precipitation record and established global or climactic intensity-duration thresholds. Landslide-triggering storms are shaded with darker blue. The panels (a)-(d) contains precipitation data for all sites while in panels (e)-(h) only verified sites are included. Points above each threshold are predicted by the threshold to be landslides, and so a larger proportion of landslides above the threshold indicates better performance.

2540x1693mm (72 x 72 DPI)

The degree of uncertainty across multiple precipitation products is evaluated for 228 landslide-triggering storm events across North America. Estimates of storm magnitudes varied by as much as 296 mm with an average range of 51 mm. Products more reliant upon ground-based observations (MRMS and NLDAS-2) were better at identifying landslides according to published intensity-duration storm thresholds (see figure). We recommend practitioners consider low-latency products and would be well-served considering more than one product as a way to confirm intense storm signals.



A Multi-sensor Evaluation of Precipitation Uncertainty for Landslide-triggering Storm Events

Elsa S. Culler*
Andrew M. Badger
J. Toby Minear
Kristy F. Tiampo
Spencer D. Zeigler
Ben Livneh

http://mc.manuscriptcentral.com/hyp