Title

Bayesian analysis of the impact of rainfall data product on simulated slope failure for North Carolina locations

Permalink
https://escholarship.org/uc/item/3m94t2vc

Journal
Computational Geosciences, 23(3)

ISSN
1420-0597

Authors
Yatheendradas, Soni
Kirschbaum, Dalia
Nearing, Grey
et al.

Publication Date
2019-06-01

DOI
10.1007/s10596-018-9804-y

Peer reviewed

Published in final edited form as:
Comput Geosci. 2019 June 15; 23(3): 495–522. doi:10.1007/s10596-018-9804-y.

Bayesian analysis of the impact of rainfall data product on simulated slope failure for North Carolina locations

Soni Yatheendradas[1,2,*], **Dalia Kirschbaum**[2], **Grey Nearing**[3], **Jasper A. Vrugt**[4,5], **Rex L. Baum**[6], **Rick Wooten**[7], **Ning Lu**[8], **Jonathan W. Godt**[6]

[1]Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20742, USA

[2]Hydrological Sciences Laboratory, NASA GSFC, Greenbelt, MD 20771, USA

[3]Department of Geological Sciences, The University of Alabama, Tuscaloosa, AL 35487, USA

[4]Department of Civil and Environmental Engineering, Henry Samueli School of Engineering, University of California, Irvine, CA 92697, USA

[5]Department of Earth System Science, University of California, Irvine, CA 92697, USA

[6]Geologic Hazards Science Center, U.S. Geological Survey, Golden, CO 80401, USA

[7]Asheville Regional Office, North Carolina Geological Survey, Swannanoa, NC 28778, USA

[8]Department of Civil & Environmental Engineering, Golden, CO 80401, USA

Abstract

In the past decades, many different approaches have been developed in the literature to quantify the load-carrying capacity and geotechnical stability (or the Factor of Safety, F_S) of variably saturated hillslopes. Much of this work has focused on a deterministic characterization of hillslope stability. Yet, simulated F_S values are subject to considerable uncertainty due to our inability to characterize accurately the soil mantle's properties (hydraulic, geotechnical and geomorphologic) and spatiotemporal variability of the moisture content of the hillslope interior. This is particularly true at larger spatial scales. Thus, uncertainty-incorporating analyses of physically based models of rain-induced landslides are rare in the literature. Such landslide modeling is typically conducted at the hillslope scale using gauge-based rainfall forcing data with rather poor spatiotemporal coverage. For regional landslide modeling, the specific advantages and/or disadvantages of gauge-only, radar-merged and satellite-based rainfall products are not clearly established. Here, we compare and evaluate the performance of the Transient Rainfall Infiltration and Grid-based Regional Slope-stability analysis (TRIGRS) model for three different rainfall products using 112 observed landslides in the period between 2004 and 2011 from the North Carolina Geological Survey database. Our study includes the Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis Version 7 (TMPA V7), the North American Land Data Assimilation System Phase 2 (NLDAS-2) analysis, and the reference 'truth' Stage IV precipitation. TRIGRS model performance was rather inferior with the use of literature values of the geotechnical parameters and soil hydraulic properties from ROSETTA using soil textural and

[*]Corresponding author: soni.yatheendradas@nasa.gov, +1 (301) 286-9135.

bulk density data from SSURGO (Soil Survey Geographic database). The performance of TRIGRS improved considerably after Bayesian estimation of the parameters with the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm using Stage IV precipitation data. Hereto, we use a likelihood function that combines binary slope failure information from landslide event and 'null' periods using multivariate frequency distribution-based metrics such as the False Discovery and False Omission Rates. Our results demonstrate that the Stage IV-inferred TRIGRS parameter distributions generalize well to TMPA and NLDAS-2 precipitation data, particularly at sites with considerably larger TMPA and NLDAS-2 rainfall amounts during landslide events than null periods. TRIGRS model performance is then rather similar for all three rainfall products. At higher elevations, however, the TMPA and NLDAS-2 precipitation volumes are insufficient and their performance with the Stage IV-derived parameter distributions indicate their inability to accurately characterize hillslope stability.

Keywords

slope stability; physically based model; sensitivity; satellite-based rain; calibration; 5 (Probability Geosciences); 7 (Hydrology)

1. INTRODUCTION

Landslides triggered by intense or prolonged rainfall impact nearly all countries and can cause extensive damage (e.g., [1, 2]). Modeling the physical response and interactions of the surface and subsurface to precipitation is central to understanding and improving landslide hazard assessment and for better anticipating the timing and location of landslides. However, current slope-stability models have primarily been limited to local or hillslope-scale investigations and study domains. This is due to limitations in the availability and quality of in situ data needed to parameterize and spatially upscale these complex models with nonlinear hydrogeotechnical relationships that effectively resolve surface and subsurface behavior (e.g., [4]). Proximate rain gauge precipitation is typically relied on as the forcing source by such modeling studies and has also reinforced the local scales of application in such studies (e.g., [3]).

Spatially distributed rain data, including from satellite, ground radar or other products with quasi-global or regional spatial coverage, provide the opportunity to apply a slope-stability model framework over larger areas. Such application scales typically involve spatially distributed model runs everywhere in the region with potentially one or more of the lateral flow connectivity components activated (e.g., overland flow, surface layer flow, subsurface layer flow). Understanding and diagnosing the model behavior for a successful application involves multiple analyses. For example, sensitivity analyses of physically based slope stability modeling outputs to deterministic rainfall input sources is done in our study.

The logical progression of such analyses that enables a systematic buildup of knowledge is to start with simple model configurations (vertical 1-dimensional or 1D) before incorporating components like lateral flow components. This enables a clearer understanding of the behavior change between such configurations. For example, in hydrology, initial 'point' 1D calibrations were conducted for sites with flux towers and soil moisture probes only (and not everywhere in the region). If landslide databases were comprehensive enough to not have missing observations, one can potentially conduct reliable analyses everywhere in a region, including at ‘null’ locations and periods. However, considering the missing observations bias existing in all databases so that a site without observed failure is not reliably null in reality, this study focuses on 1D simulation of reliable landslide-observed sites only.

Few studies have examined the sensitivity of slope stability models to satellite-based or other precipitation sources. Studies such as those of [5] have successfully applied satellite rain directly within a spatially distributed and laterally connected physically based model, but their results involved a large number of false alarms within the modeling framework (i.e., the model predicted unstable pixels without observed landslides). Our work considers the Stage IV multi-sensor product available over the continental United States as the ‘truth’ rain data [6]. We then explore the general sensitivity of stochastic landslide predictions to deterministic rainfall products including the Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis Product (TMPA) [7, 8] and the North American Land Data Assimilation System Phase 2 (NLDAS-2) analysis [9, 10].

Ensemble-based (or uncertainty-incorporating) analysis in dynamic slope-stability modeling has typically been conducted on single or few slopes [11, 12] or synthetic ones (e.g., [13]), and rarely on multiple real natural slopes or at the regional/watershed scale [14, 15]. Many of these studies have used rainfall forcing that is synthetically constructed from long-term rainfall characteristics (e.g., intensity, duration) [11, 14, 15]. To approach slope-stability modeling over larger regions, a large-sample approach is critical. For example, in hydrologic research, recent work has considered a large number of catchments [16]. This moves away from the earlier ‘depth’ focus of intensive investigation at a few heavily instrumented locations or catchments towards the ‘breadth’ offered by a large sample in order to balance both. Such a balancing approach provides the necessary detail and robustness required for process understanding across larger spatial domains to facilitate better transferability of model parameters across regions. Our study models local slope stability at multiple locations across North Carolina to better resolve the effect of regional heterogeneity in precipitation and landscape properties.

The overarching question addressed in this study is: How do modeled landslide responses change with rain data source? We attempt to answer this question using the Transient Rainfall Infiltration and Grid-based Regional Slope-stability (TRIGRS) model [4]. We evaluate the performance of TRIGRS for three different rainfall products: Stage IV, TRMM, and NLDAS-2 using default (*a priori*) and calibrated (e.g., *a posteriori*) values of the model parameters. Note, the *default* parameterization (also referred to as uncalibrated parameter values in this study) may depend upon (among others) field expertise, earlier publications, and/or pedotransfer functions. Thus, the default or uncalibrated parameter values may express some form and/or level of calibration.

When the default model parameterization is inadequate, calibration may be necessary to improve model performance and enhance the consistency and reliability of the simulated output (e.g., [12]). Parameters of slope-stability models such as TRIGRS are typically calibrated manually. An alternative to this is exhaustive sampling, an example of which is the study of [74] who used Generalized Likelihood Uncertainty Estimation (GLUE: [75]) to derive so-called behavioral parameter values and quantify model predictive uncertainty. In that paper, the prior range of each of the four areal-constant TRIGRS model parameters was discretized into 10 equidistant values. From this ensemble of 10,000 parameter vectors, the authors then focused their attention to those 25 samples which most accurately described the observed data, while minimizing the distance to the perfect classification (D2PC: [76, 77]) on the receiver operating characteristic (ROC) simulated by each parameter vector ([78]).

Our study uses Bayesian inference coupled with Markov chain Monte Carlo (MCMC) simulation to quantify parameter and predictive uncertainty of a physically based landslide model. This demands CPU-intensive numerical simulation of hillslope stability for many different parameter values and places a premium on computational resources and MCMC search efficiency. This may explain the lack of landslide model calibration studies published in the literature (e.g., [12]). Thus, our research addresses the following question: Does TRIGRS model calibration substantially improve slope failure simulation with respect to the default parameterization? The answer to this question should be of interest to modelers and/or practitioners, and provokes a follow-up question: What changes in the model parameter distributions lead to successful slope-stability simulations? Lastly, we investigate the effect of elevation on TRIGRS model performance. This research may be summarized with a third and final question: How does elevation affect modeled slope failure for each of the three rainfall products used herein?

The remainder of this paper is organized as follows. Section 2 describes the study data: the landslide events, rainfall, monitoring locations and soil properties. Section 3 introduces the models and methodologies used in this study for characterizing uncertainty in landslide simulations and sensitivity to the rainfall forcing data source. Results are presented in Section 4, and this is followed by Section 5, which is a discussion of our main findings and conclusions.

2. DATA: LANDSLIDES, RAINFALL, SOILS AND MONITORING LOCATIONS

2.1 The NCGS landslide geodatabase

The North Carolina Geological Survey (NCGS) landslide/slope movement geodatabase version dated April 16, 2015, has landslide data from 1916 to the present; however, just about 12% of these landslides are dated post-2000. In addition to the geographic location and approximate timing of the landslides, the geodatabase also includes information on estimated soil depth and properties for some landslides; however, these values can further vary within the initiation zone of an individual landslide. We did not consider landslides where field notes indicated that the slope was anthropogenically altered or disturbed (e.g., cut slope, road cut, embankment), had slope movement material classified as rock, or were not field-visited/verified. Fig. 1 shows the 112 landslides which satisfy listed requirements for the present study, out of which 94 (84%) were triggered by Hurricanes Frances and Ivan that occurred only 1 week apart in September 2004 [17], with the remaining events happening in July 2011.

The landslides have associated spatial location uncertainty radii based on their source (50 ft for lidar-based field verification, 40 ft radius for GPS, and 400 ft for 7.5’ quad), and we assigned a 20 ft locational uncertainty corresponding to the LiDAR resolution for orthophotography source. This uncertainty was used to buffer the locations and intersect with soil polygons from the Soil Survey Geographic database (SSURGO: [18]) version 2.3.2, allowing a possibility of potentially multiple SSURGO soils per landslide location. We obtained a total of 180 ‘sites’ or combinations of landslides and associated possible SSURGO soils that were then subject to physically based slope-stability simulations. Each site is then actually a polygon but modeled as a single pixel domain with the geomorphologic uncertainties of elevation and slope included among the model parameter uncertainties considered in this study.

2.2 Rainfall data

The Stage IV rain analysis [6, 19] is provided by NCAR/EOL under sponsorship of the National Science Foundation and is based on the multi-sensor ‘Stage III’ analyses (on local 4 km polar-stereographic grids) produced by the 12 River Forecast Centers (RFCs) in the continental United States (CONUS). The Stage III data are already a combination of gauge observations and radar-calculated reflectivity and includes a manual Quality Control process. The National Centers for Environmental Prediction (NCEP) mosaic the Stage III analysis into the Stage IV product that spans the CONUS spatial domain.

For our satellite precipitation product, we resort to version 7.0 of the Tropical Rainfall Measurement Mission Multi-Satellite Precipitation Analysis, or TMPA V7 [8, 20]. TMPA, also referred to as V7, merges passive microwave, active radar, and infrared observations from multiple different satellites to create a quasi-global (±50° latitude) precipitation dataset with 3-hourly rainfall estimates on a grid with spatial resolution of 0.25 degrees. We assume the 3-hourly rain rates to hold constant over their associated 3-hour windows and derive the hourly rainfall estimates for this study by simply assigning or averaging over the modeled 1-hourly time intervals as relevant. In practice, there may be variability within the sampling time window due to the overpass frequency of the constellation of satellites used to develop this product; however, nearest neighbor interpolation provides the most reliable estimates for hourly intervals. The V7 product is available with a latency of about two months and includes a monthly rain gauge-based bias correction with temporal coverage from 1998-present.

As third and last precipitation product, we use the example ‘File A’ forcing dataset from Phase 2 of NASA’s North American Land Data Assimilation System (NLDAS-2: [9, 10]). This rainfall product exhibits a spatial resolution of 0.125 degrees and disaggregates to hourly data the daily CPC-Unified gauge-only analysis [21, 22] (before 2012) or the operational CPC product (after 2012). Both these CPC products use a statistical topographic correction based on the PRISM climatology by [23]. Temporal disaggregation depends on the availability of the following rainfall data sources listed in order of decreasing importance: (i) hourly WSR-88D Stage-II Doppler radar-based precipitation estimates from 1996-present, (ii) half-hourly 8-km CMORPH hourly precipitation analyses [24] from 2002-present, (iii) CPC CONUS/Mexico gauge-based Hourly Precipitation Data (HPD: [25]), and

(iv) 3-hourly North American Regional Reanalysis Regional Climate Data Assimilation System precipitation (NARR/R-CDAS: [26]). Note that this NLDAS-2 rainfall product exhibits a coarser spatial resolution than the Stage IV precipitation data.

2.3 Monitoring locations

Fig. 1 shows the four in situ monitoring locations where hydraulic and geotechnical properties including shear strength were estimated through laboratory testing. The hydraulic properties exhibit hysteresis and were therefore derived during wetting and drying conditions from laboratory experiments conducted on the soil specimens. Soil properties available at these locations and used in this study are the porosity, $\Phi_{\mathrm{m}}(-)$, specific gravity, $G_{\mathrm{S,m}}$ (−), cohesion for effective stress, c'_{m} (Pa), the friction angle for effective stress, $\phi'_{\mathrm{m}}(°)$, the reciprocal of the air-entry pressure head, α_{m} (m^{-1}), the saturated and residual soil moisture contents, $\theta_{\mathrm{s,m}}$ and $\theta_{\mathrm{r,m}}$ ($\mathrm{m}^3/\mathrm{m}^3$), respectively, and $K_{\mathrm{s,m}}$ (m/s), the saturated soil hydraulic conductivity, where the subscript ‘m’ signifies monitoring location. The porosity and specific gravity were used to compute the void ratio, e_{m} (−), and unit weight of soil, $\gamma_{\mathrm{s,m}}$ ($\mathrm{kg/m^2s^2}$) required by TRIGRS for landslide simulation:

$$e_{\mathrm{m}} = \Phi_{\mathrm{m}}/(1 - \Phi_{\mathrm{m}}) \tag{1}$$

$$\gamma_{\mathrm{s,m}} = \gamma_{\mathrm{w}}(G_{\mathrm{S,m}} + e_{\mathrm{m}})/(1 + e_{\mathrm{m}}), \tag{2}$$

where γ_{w} ($\mathrm{kg/m^2s^2}$) signifies the unit weight of water.

2.4 SSURGO soils data

We use SSURGO version 2.3.2 [18] for the soil hydraulic parameters, since its level of spatial discretization over North Carolina assigns unique soils among the monitoring locations when compared against other soil maps like STATSGO and FAO. SSURGO has map units that are delineated polygons and correspond to soil materials that are either consociations named for the dominant soil taxon with other similar soils and dissimilar inclusions present, or they can be complexes and associations that have dissimilar soil components with possibly a dominant soil component. Miscellaneous land types or areas of water may also be included. SSURGO Map units are denoted by ‘Musyms’ (Map Unit Symbols).

We note that Musyms with identical first two letters differ only in their slope range values. For example, the Musym values of *EdB, EdC, EdD, EdE* and *EdF* have the first two letters ‘*Ed*’ denoting a map unit of stony Edneyville-Chestnut complex, while the last letter denotes different contiguous slope ranges where the slopes increase when going from *EdB* (2 to 8 percent slopes) to *EdF* (50 to 95 percent slopes). We group the sites according to these first two letters of Musym that we then use and refer to as Musym soils to denote such relevant soil groups (e.g., *Ed*), unless explicitly stated otherwise. Table 1 lists the Musym soils and their descriptive names/components for our landslide sites and monitoring locations. Note that these descriptive names/components correspond to the dominant soil(s) at the study sites considered, since they can spatially vary within a specific Musym soil.

This study uses SSURGO-based available information on the percentages of sand, silt, clay, and bulk densities at each landslide site's vertical soil profile. The Musym soils at the monitoring locations 'Poplar Cove 2,' 'Mooney Gap 4,' 'Mooney Gap 1,' and 'Bent Creek 1,' respectively, are *Ed* (Edneyville-Chestnut complex, stony), *Cu* (Cullasaja-Tuckasegee complex, stony), *Cp* (Cleveland-Chestnut-Rock outcrop complex, windswept) and *Pw* (Plott fine sandy loam, stony).

3. METHODOLOGY AND MODELS

3.1 The TRIGRS slope-stability model

We use version 2.1 of the event-based TRIGRS model [4, 27, 73] to analyze, explore and evaluate the usefulness of the different rainfall data products in characterizing hillslope stability for our test sites in North Carolina. This model, developed by the USGS, solves numerically for the stability of the hillslope using a single integrated measure of its load-carrying capacity. This measure, coined the Factor of Safety (F_S), depends strongly on precipitation and the hydraulic and geotechnical properties of the hillslope interior (e.g., soil mantle and underlying bedrock). Infiltrating water elevates pore pressures within the soil mantle, which in turn reduces the shear strength. TRIGRS-simulated traces of F_S summarize hillslope stability during the rainfall event and/or subsequent dry period and can be used to pinpoint the timing and initiation of shallow, rainfall-induced landslides. Slope failure occurs when the F_S value becomes less than unity.

TRIGRS solves for the pore water pressure in the variably saturated hillslope using a linearized version of the one-dimensional Richards' equation [28]. This approach assumes an isotropic and homogeneous soil mantle. The load-carrying capacity of the hillslope is then characterized using one-dimensional, infinite-slope stability analysis [29]. In Taylor's analysis, slope stability is characterized by the ratio of resisting basal Coulomb friction to gravitationally induced downslope basal driving stress. This unitless ratio is F_S and depends on time, t (days), and the depth, Z (m), of the soil mantle as follows

$$F_{\mathrm{s}}(Z,t)=\frac{\tan(\phi')}{\tan(\delta)}+\frac{c'-\chi(Z,t)\psi(Z,t)\gamma_{\mathrm{w}}\tan(\phi')}{\gamma_{\mathrm{s}}Z\sin(\delta)\cos(\delta)}, \tag{3}$$

where ϕ' (°) denotes the soil friction angle for effective stress, c' (Pa) signifies the soil cohesion for effective stress, χ (−) is the degree of saturation, $\psi(Z,t)$ (m) characterizes the soil water pressure head at depth Z, γ_w and γ_S are the unit weights of water and soil in kg/m^{2}s^2, respectively, and δ (°) represents the slope angle. The degree of saturation, χ, is computed as follows [30]:

$$\chi(Z,t)=\frac{\theta(Z,t)-\theta_{\mathrm{r}}}{\theta_{\mathrm{s}}-\theta_{\mathrm{r}}}, \tag{4}$$

where $\theta(Z, t)$ (m^3/m^3) signifies the volumetric moisture content in the soil mantle at time t and depth Z. Note, that the product $\chi(Z,t)\psi(Z,t)\gamma_{\mathrm{w}}$ in the numerator of Equation (3) equates to the suction stress, σ_S, in units of Pascal.

We execute TRIGRS in a mode commensurate with hydrologic modeling and simulate one-dimensional soil moisture flow and storage in a variably saturated finite soil domain bounded below by bedrock. We compute the Factor of Safety in Eq. 3. for each depth, Z, in the soil mantle and use the minimum value in the profile as our measure of hillslope stability at time t. Thus, for each site in our region, we simulate traces of F_{S} during multiple different durations, classified as null or landslide periods.

The original TRIGRS code creates many different output files (one separate file for each timestep and model output) which demands significant computational time and resources for multi-core distributed simulation as used in the present study. Hence, we recoded TRIGRS to write the simulated outputs of all timesteps into a single file. TRIGRS has 12 adjustable model parameters and corresponding distributions (see the first 12 parameters in Table 2, explained more in following Section 3.3).

3.2 Likelihood function for performance evaluation

The accuracy and reliability of the simulated output of the TRIGRS model can be determined by comparison against measured data. Since hillslope interiors are difficult to access, characterize, and monitor in situ, field experimentation is often impractical, time-consuming, labor-intensive and expensive. Observations of hillslope stability are, therefore, often limited to binary data on whether slope failure and/or mass movement was observed or not during the period of interest. Such categorical information, or soft data, from geotechnical experts and/or field experimentalists is not always easy to use for model evaluation and/or parameter estimation. As a consequence, geotechnical studies often resort to spatiotemporal variations in the simulated Factor of Safety, F_{S}, to obtain time-varying probabilities of slope failure during landslide events (e.g., [12, 15]). The so-obtained probabilities may be of practical value and/or use but as they depend only on simulated output they cannot be used to quantify TRIGRS model performance for individual ‘point’ sites nor help determine behavioral (posterior) values of its slope-stability parameters. This necessitates the use of slope stability observations and a likelihood function that measures in a probabilistic sense the accuracy of the TRIGRS-simulated output. Unfortunately, the formulation of such likelihood function is not particularly easy in the present context with binary data on landslide occurrence. Maximum likelihood estimation procedures for categorical variables developed in probability theory and statistics, will not suffice in the present situation with innate underlying ordering of the null and landslide periods. Indeed, we desire that the TRIGRS model predicts accurately slope failure probability and timing. Thus, we must first postulate a function that adequately characterizes the likelihood of simulated landslide occurrences over our spatiotemporal domain of interest. Insofar as possible, this function should satisfy first-order statistical principles.

For each of the sites in our database, the likelihood function quantifies TRIGRS model performance for several different rainfall periods which are either classified in our database as a null (no mass movement) or landslide period (slope failure). For slope failure sites, the null periods always precede the landslide period. Due to limitations in the inventory, it is possible that for some failure sites the database has erroneously classified a rainfall event as a null period, and thus, slope failure and mass movement did take place in one or more of the storm events leading up the landslide period. However, the probability of slope failure having occurred during those null periods at a slope failure site and being unnoticed during the recording into the database of the landslide period at that site is likely very low. Hence, we consider the so-called null periods to be free of landslide activity. Therefore, the binary classification used herein should adequately portray the stability of our hillslopes during each of the successive rainfall events and provide robust TRIGRS parameter estimates and modeling results.

Our likelihood function combines binary information of slope failure occurrence or non-occurrence from both the landslide event period and earlier ‘null’ periods. This function merges into one statistical measure of model performance the False Omission Rate, R_{fo} (−), the ratio of wrongly simulated stable slopes or non-failures to the total number of simulated stable slopes, and the False Discovery Rate, R_{fd} (−), the ratio of wrongly simulated failures to total number of simulated failures. These two unitless metrics are computed as follows

$$R_{\mathrm{fo}} = n_{\mathrm{FN}}/(n_{\mathrm{FN}} + n_{\mathrm{TN}}) \tag{5}$$

$$R_{\mathrm{fd}} = n_{\mathrm{FP}}/(n_{\mathrm{FP}} + n_{\mathrm{TP}}), \tag{6}$$

and are based on a 2×2 contingency table or confusion matrix where n_{FN}, n_{TN}, n_{FP}, and n_{TP} denote the long-term number of false negatives, true negatives, false positives and true positives, respectively. Here long term denotes the scenario in which TRIGRS might have been run for an infinitely or sufficiently long simulation duration comprising of both null and/or landslide periods. The values of R_{fo} and R_{fd} range between 0 (perfect performance) and 1 (poor performance), thus lower values of both metrics are preferred. Thus, if for some site, TRIGRS had simulated a stable slope during a period in which a landslide was observed, then that period is classified as a false negative, and a value of unity is added to the long-term n_{FN} of this site. Now, the total number of simulated instances with a stable slope is equivalent to $n_{\mathrm{FN}} + n_{\mathrm{TN}}$. The likelihood of simulating a stable slope for that period is the long-term value of $n_{\mathrm{FN}}/(n_{\mathrm{FN}} + n_{\mathrm{TN}})$, which, by definition, is equal to R_{fo} in Eq. (5). If, instead, TRIGRS simulated slope failure during a landslide period then this is considered one occurrence of the true positive n_{TP}. The likelihood of simulating slope failure for that same period is the long-term $n_{\mathrm{TP}}/(n_{\mathrm{FP}} + n_{\mathrm{TP}})$, or $(1 — R_{\mathrm{fd}})$, with R_{fd} computed in Eq. 6.

If **x** is the vector of TRIGRS parameter values and $\tilde{\mathbf{y}}$ our vector of observations (i.e., slope stability for null periods and slope failure for landslide periods), then our likelihood function, $L(\mathbf{x} \mid \tilde{\mathbf{y}})$, aggregates into one statistical measure of model performance the False Omission and False Discovery Rates, R_{fo} and R_{fd}, from Eq. 5 and 6 respectively, for the null and landslide events periods as follows:

$$L(\mathbf{x} \mid \tilde{\mathbf{y}}) = \left[(1 - R_{\mathrm{fo}})^{N_{\mathrm{n},s}} R_{\mathrm{fd}}^{N_{\mathrm{n,f}}}\right]^{\left(\frac{1}{N_{\mathrm{n}}}\right)} \left[(1 - R_{\mathrm{fd}})^{N_{\mathrm{l,f}}} R_{\mathrm{fo}}^{N_{\mathrm{l},s}}\right]^{\left(\frac{1}{N_{\mathrm{l}}}\right)}, \tag{7}$$

where N_{n} (N_{l}) denotes the number of actual number of null (landslide) periods that we simulate, $N_{\mathrm{n,f}}$ and $N_{\mathrm{l,f}}$ count the number of simulated slope failures during these null and landslide periods, respectively, $N_{\mathrm{n,s}} = N_{\mathrm{n}} - N_{\mathrm{n,f}}$ and $N_{\mathrm{l,s}} = N_{\mathrm{l}} - N_{\mathrm{l,f}}$. Thus, $N_{\mathrm{n,s}}$ and $N_{\mathrm{l,s}}$ count the number of stable slopes simulated during the null and landslide periods, respectively.

The exponents, $1/N_{\mathrm{n}}$ and $1/N_{\mathrm{l}}$, which act on the two terms in the square braces give a similar weight in the likelihood function to the null and landslide period groupings, respectively. This weighting addresses the class imbalance problem [86] frequently encountered in classification problems in machine learning and parameter estimation. In the absence of such weighting, imbalance between class frequencies will result in a low sensitivity of the likelihood function to the infrequent class (landslide periods in our study) and favor heavily the fitting of the frequent class (null periods in our study). Thus, our exponents $1/N_{\mathrm{n}}$ and $1/N_{\mathrm{l}}$ promote inference of parameter values that balance the performance of TRIGRS during the null and landslide periods. The $N_{\mathrm{n}} > N_{\mathrm{l}}$ in our study reflects the general long-term inequality of null instances being much more preponderant than landslide periods. For each slope failure in our database, we consider five null periods, $N_{\mathrm{n}} = 5$, and $N_{\mathrm{l}} = 1$. We leave for future studies the investigation of the effect of N_n on the behavioral parameter values and simulation results.

The use of Eq. 7 has one important drawback. Parameter vectors which consistently simulate slope failure (or non-failure) for all null and landslide periods are undesirable since they provide no information for the search algorithm to discriminate between the two groupings. Such ‘nonbehavioral’ solutions are unproductive and should be discarded early on to promote convergence of the DREAM$_{(\mathrm{zs})}$ algorithm to the appropriate behavioral solution space. In our study, we therefore assign these aberrant parameter vectors a likelihood of zero. The modified likelihood function, $L_{\mathrm{m}}(\mathbf{x} \mid \widetilde{\mathbf{y}})$, now reads as follows:

$$L_{\mathrm{m}}(\mathbf{x} \mid \widetilde{\mathbf{y}}) = \begin{cases} 0 & \text{if } N_{\mathrm{n,f}} + N_{\mathrm{l,f}} = 0 \text{ or } N_{\mathrm{n,f}} + N_{\mathrm{l,f}} = N_{\mathrm{n}} + 1 \\ L(\mathbf{x} \mid \mathbf{y}) & \text{otherwise.} \end{cases} \tag{8}$$

This modified likelihood as an improvement over Eq. 7 may prove futile for some sites. If the prior parameter space is largely made up of nonbehavioral solutions, then it may be difficult to determine a suitable search direction as the likelihood (and the gradient thereof) will be zero. This complicates unnecessarily the search for behavioral solutions. We will revisit this issue in the remainder of this paper.

The NCGS geodatabase does not report specific (instantaneous) time values for each landslide event, but rather provides a failure time range on the order of days. For our study area, this failure period can be as short as a day (e.g., September 17, 2004) to as long as 12 days (e.g., September 6–17, 2004). Therefore, a TRIGRS simulation is deemed successful if slope failure is modeled within the recorded time period of each landslide event. For a null period, however, we consider slope failure to be simulated if it occurs at any time within the null period. For each site and rainfall data source used in this study, the start and end time of each TRIGRS simulation was determined manually for each null and landslide period. Simulations for each site are initiated at the onset of rainfall and terminated at the end of each storm event when precipitation has ceased. If we denote with t_{b} and t_{e} the beginning and end date of the storm (rainfall) event in days, then for some of the sites the corresponding simulation period, $\Delta t = t_e - t_b$, may not be long enough to warrant an accurate description of hillslope stability. Indeed, pore water pressures may continue to rise into the dry hours after completion of a storm event. If deemed appropriate, we, therefore, changed the simulation end time to $t_e + 0.25\Delta t$. This extension of the simulation period with $0.25\Delta t$ days should give sufficient opportunity for the rainfall to infiltrate into the soil mantle and to accurately portray failure simulation in the dry days immediately following a storm event. For some of the sites, the so-obtained time interval of simulation did not encapsulate the period of recorded slope failure in the NCGS database. In those cases, the length of the simulation period was enhanced further by moving forward t_b and/or pushing back to a later time t_e.

3.3 Uncalibrated parameter distributions

In general, the information in the NCGS database corresponds with the thicker exposures of the soil in the initiation zone (IZ) for each landslide. As the soil depth (and properties) within this IZ may not be constant, this begs characterization of parameter and/or model output uncertainty. Table 2 details information about the *a priori* (or uncalibrated) parameter distributions used in the present study and described in the following subsections.

3.3.1 Hydraulic parameters—TRIGRS does not characterize explicitly the complex stratigraphy of hillslope interiors other than an impermeable or leaky lower boundary with constant representative values for the parameters of the soil mantle. We use SSURGO to derive estimates of the texture, bulk density and soil depth at each site in our study domain. If SSURGO estimates of the soil depth are smaller than measured depths to bedrock at a failure site, we use the entire SSURGO soil profile to obtain representative values of the TRIGRS soil properties. Instead, if the soil depth estimates derived from SSURGO exceed the bedrock depth, a slicing procedure is used that maps the relevant SSURGO soil properties to the soil mantle at our failure site.

To determine the relevant soil thickness slice, we consider the controls on the hydrological response to be near the failure depth that is usually near-coincident with the bedrock/soil interface. This means that the timing of pore water pressure rise and associated failure will be largely controlled by the hydrological properties of the SSURGO layer (horizon) immediately overlying the bedrock. SSURGO layers further above will have some (lesser) influence on the response with their effect attenuating with distance. This also reflects the topmost layer being much more permeable than anything below and making a negligible contribution to the time delay between rainfall and pore pressure rise at the depth of failure. Hence, we consider the representative SSURGO soil thickness slice (possibly spanning multiple layers) to be immediately above the level of the bedrock/soil interface. For a failure site where the SSURGO Musym soil has multiple areal components, each component will have its own soil thickness slice. Therefore, for failure sites, the representative soil textural percentages and bulk density are calculated by weighted averaging with the relevant layer thicknesses as the weights. The overall representative values of the textural percentages and bulk density for the Musym soil are then obtained by averaging over those components with the areas as the weights.

The soil textural (percentages of sand, silt and clay) and bulk density of the hillslope interior are input to ROSETTA [31, 32] to obtain uncalibrated distributions for the hydraulic parameters θ_r (residual soil moisture content), θ_s (saturated soil moisture content), α (inverse of the air-entry value), and K_s (saturated soil hydraulic conductivity). Percentages of sand, silt, and clay are a minimum input requirement for ROSETTA and hydraulic parameters are further constrained when bulk density and porosity data are also available. ROSETTA provides lognormal distributions for K_s (e.g., [14, 33, 34]) and α (e.g., [12]), and normal distributions for θ_r and θ_s. We do not assign a lognormal distribution to the hydraulic diffusivity (e.g., [14, 33]), but rather follow the TRIGRS equations and compute the diffusivity as a ratio of K_s and the specific storage, S_s. Based on field measurements in active landslides [4, 35, 36], we assign a uniform distribution to S_s with lower and upper bounds of 0.005 and 0.5 m^{-1}, respectively. The result is an intrinsic correlation between the computed diffusivities and sampled K_s values, which does not have to be specified a priori by the user (e.g., [37]).

If we ignore correlation among the hydraulic parameters of the soil mantle, then this may result in unrealistically large prior probabilities and undesirable or physically unrealistic soil water retention and hydraulic conductivity functions [38]. We use the Monte Carlo approach of [38, 12] to derive the correlation (covariance) matrices of the hydraulic parameters. For each site, we create 250 different realizations of sand, silt and clay percentages by drawing at random from a three-variate normal distribution. The means of this distribution are the SSURGO-derived percentages of sand, silt and clay, while the covariance matrix has values of 0.25 on the main diagonal and off-diagonal elements of −0.125. Each realization is normalized so that the fractions of sand, silt and clay add up to 100%. This ensemble of soil texture realizations is then used as input to ROSETTA and produces a matrix of 250 × 4 having corresponding values of the hydraulic parameters θ_r, θ_s, α, and K_s of the soil mantle. Log-transformed values are used for α and K_s. This matrix is then used to compute pairwise correlation coefficients of θ_r, θ_s, α, and K_s.

If the Musym soil of the failure site coincides with that of the monitoring location (i.e., *Ed, Cu, Cp*, or *Pw*), the uncalibrated distributions of θ_r, θ_s, α, and K_s derived from ROSETTA are refined further using soil hydraulic data from the monitoring location. We present here a recipe of this procedure for some property x at some measurement site. Per ROSETTA, x is supposed to be μ_0-mean normally distributed with variance σ_0^2. At this location we have available n different "observations" of x stored in the vector, $\tilde{\mathbf{x}} = \{\tilde{x}_1, \dots, \tilde{x}_n\}$. If we treat each observation to be the mean of another distribution with variance σ_0^2, then we can refine our ROSETTA-derived estimates of μ_0 and σ_0^2 for soil property x at the measurement site as follows, $\mu_{0,new} = \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} \tilde{x}_i}{\sigma_0^2}\right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_0^2}\right)$ and $\sigma_{0,new}^2 = 1 / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_0^2}\right)$. This simplifies to $\mu_{0,new} = \left(\mu_0 + \sum_{i=1}^{n} \tilde{x}_\mathrm{i}\right)/(1+n)$ and $\sigma_{0,new}^2 = \sigma_0^2/(1+n)$. Among the n observations of each site, we include values of the hydraulic parameters of the soil mantle derived during wetting and drying experiments. Section 3.3.6 below details our treatment of the multivariate dependencies among the hydraulic parameters of the hillslope.

3.3.2 Geotechnical parameters—Whereas pedotransfer functions can be used to derive probability distributions of the hydraulic parameters of the hillslope interior based on (among others) stipulated variations in the texture of the soil mantle (see previous section), it is not particularly easy to a priori determine suitable values of the geotechnical parameters, let alone their underlying multivariate probability distribution. We use the following approach to determine the a priori (or uncalibrated) distributions of the geotechnical parameters. First, we use the textural information (percentages of sand, silt, and clay) from SSURGO to classify soil type (e.g., sandy loam, silt) using the textural triangle [39]. Then, for each soil type, we assign ranges and distributions of the geotechnical parameters. For the unit weight of soil, γ_s, we use the ranges listed in the NAVFAC DM 7.01 manual [40] and found online at http://www.geotechnicalinfo.com. For the effective cohesion, c', and effective friction angle, ϕ', we use the ranges reported online at http://www.geotechdata.info (see Table 3 for the ranges for the soil textural classes at the landslide failure sites). We use normal distributions to describe the uncertainty of γ_s (e.g., [14, 33]), c' and ϕ' (e.g., [41–44]). We further assume that c' and ϕ' are independent [45–49]. This latter assumption is practical and convenient but may not always be appropriate [50]. If the Musym soil of the failure site coincides with that of the monitoring location, then we use available data on γ_s, c' and ϕ' at that location to refine the geotechnical parameter distributions using the merging procedure outlined in the previous section. Thus, our approach links the geotechnical parameters to the properties of the soil mantle, and honors spatial variations in soil type across our study region. This should be an improvement over the use of spatially constant geotechnical properties over all soil types (e.g., [15]).

3.3.3 Geomorphologic parameters—We assume that the slope, δ, and the depth to bedrock, $Z_{\max}$ (e.g., [14,33]), are well described with a normal distribution. Multiple measurements of the undulating ground slope with a handheld clinometer or Brunton compass were averaged to obtain the value of the slope, δ, recorded in the NCGS database. The undulation of up to 3° (± 6° around the mean slope) was used to construct the 95% ranges. For sites without reported slope measurements in the NCGS database, we use the 20 ft horizontal resolution LiDAR digital elevation models (DEMs) made available with elevations rounded to the nearest foot (https://services.nconemap.gov/). The Root Mean Square Error (RMSE) of the so-derived spatially coincident LiDAR slopes was found to be 4.31°. This RMSE was derived by comparing the LiDAR slopes against sites reported in the NCGS database. We assume this RMSE to be an unbiased estimator of the standard deviation of the LiDAR-derived slopes. Thus, for sites where the NSGC database does not provide slope values, we construct the 95% ranges of δ by centering ± 8.62° (or twice of 4.31°) on the LiDAR-derived coincident slopes.

Values of $Z_{\max}$ were derived from listed failure depths, or Z_r values, in the NCGS database. Based on field measurement guidelines for failure sites with coincident available Z_r values, we set the standard deviation to 0.5 ft or 1 ft respectively for Z_r values below or above 10 ft. To obtain $Z_{\max}$ distributions at sites where failure depth measurements are not readily available, we resort to a simple regression function of Z_r against slope from the data at sites with available Z_r measurements. This is consistent with previous studies where terrain and bedrock properties and/or climatology have been used as predictors of the regolith depth

[51–56]. However, such statistical regression functions of one region (e.g., North Carolina) often lack the underlying physical rigor to make them applicable in other regions. For sites with no Z_r measurements but with available δ values in the database, we obtained the following cubic regression equation [52]:

$$Z_r = (1.4127 - 0.0061\delta)^3, \tag{9}$$

with standard deviation of Z_r of about 3.948 m. For all other locations without reported values of Z_r and δ in the database, we use a cubic equation with coefficients derived from fitting against the LiDAR-derived δ values:

$$Z_r = (1.7409 - 0.0142\delta)^3 \tag{10}$$

The standard deviation of the so-derived values of Z_r amounts to 3.942 m.

Based on field expertise, we obtain the mean of the Z_{max} distribution by adding an allowance to Z_r. This allowance is taken as the minimum of 1 foot and 10% of Z_r (the former was the field-reported maximum of the observed differences between Z_{max} and Z_r for sites in North Carolina). The standard deviation of Z_{max} is derived by multiplying the standard deviation of Z_r with Z_{max}/Z_r.

3.3.4 Initial conditions—The range of valid initial depths to the water table, d_{init}, are those that do not give slope failure at the start of the slope-stability simulation but potentially can at a later time during the failure-observed event simulation depending on the rain. For any d_{init} value, the pressure head (ψ) increases with depth (Z) for $Z > d_{init}$ until a maximum at depth Z_{max} so that the tendency for slope failure tends to be highest and F_s tends to be lowest at or near Z_{max}. Considering an increase in d_{init}, the F_s at Z_{max} increases from a minimum when $d_{init} = 0$ so that the $F_s = 1$ failure threshold is reached for some $d_{init} > 0$, and we use this threshold d_{init} (i.e., $d_{init,\, F_S = 1}$) as the lower bound for the valid range of d_{init} mentioned above. This valid range can potentially be a very small fraction of the depth to bedrock, Z_{max}. Hence, for any combination of the δ, c', ϕ' and parameters, the $d_{init,\, F_S = 1}$ value is calculated further below from Eq. 3, constituting an extrinsic uncalibrated parameter dependency in this study.

Now, Eq. 3 is actually an addition of the following 3 terms F_f, F_c and F_w:

$$F_f = \frac{\tan(\phi')}{\tan(\delta)} \tag{11}$$

$$F_c = \frac{c'}{\gamma_s Z_{max} \sin(\delta)\cos(\delta)} \tag{12}$$

$$F_W = -\frac{\chi\psi\gamma_w \tan(\phi')}{\gamma_s Z_{max} \sin(\delta)\cos(\delta)} \tag{13}$$

This means that the upper bound on F_s will always be at the maximum magnitude of the negative suction stress (and maximum positive value $F_{w,\max}$ of F_w) near the basal boundary $Z = Z_{\max}$ during conditions when the soil profile is dry. To obtain $F_{w,\max}$, we note that the online charts for the monitoring sites in North Carolina show suctions ranging from −5 to −80 kPa. These suctions correspond to soil water pressure head values between −0.5 and −8.2 m. For each ψ value in this range, we use the following equation [4]:

$$\theta = \theta_r + (\theta_s - \theta_r)\exp(\alpha\psi), \tag{14}$$

to calculate the corresponding volumetric moisture content of the soil profile. Note that the exponent in the above equation should actually use the entity ψ^* rather than ψ, where = ψ^* − ψ_0 [4]. Yet, as we use $\psi_0 = 0$, it suffices to use ψ in the exponent of Eq. 14. This exponent must be smaller than zero to ensure that $\theta_r < \theta < \theta_s$. Thus, for the general case of a nonzero ψ_0, we must subject ψ^* to another constraint so that its value cannot exceed zero.

Now that the value of the volumetric moisture content is known, we next use Eq. 4 to compute the degree of saturation of the soil profile. As the moisture content and saturation degree of the profile decrease with increasing magnitude of the soil water pressure head, it is difficult to determine a priori which value of $\psi \in [-8.2, -0.5]$ maximizes the value of F_W in Eq. 13. Therefore, we discretize the interval of the soil water pressure head in equidistant steps of 0.1 m and determine which product of $\chi\psi$ in the numerator of Eq. 13 maximizes the value of F_W, that is, $F_{w,\max}$. We then combine Eq. 3 and Eqs. 11–13 to yield the following identity:

$$F_{s,\mathrm{upper}} = F_c + F_f + F_{w,\max} \tag{15}$$

We derive the lower bound of F_s using the following built-in TRIGRS equation in the saturated soil zone:

$$\psi = \beta h = \beta(Z_{\max} - d_{\mathrm{init}}) \tag{16}$$

where the β factor converts the height of water, h, above depth $Z_{\max}$ to its corresponding pressure head value. If $d_{\mathrm{init}} = 0$, and thus the entire soil profile is saturated with water, then the maximum value of ψ is found at $Z = Z_{\max}$ and equates to $\psi_{\max} = \beta Z_{\max}$. The F_s value at the bottom of the soil profile acts as our constraint at the lower boundary of the soil column:

$$F_{s,\,Z_{\max},\,d\,=\,0} = \frac{\tan(\phi')}{\tan(\delta)} + \frac{c' - (\beta z_{\max})\gamma_w\tan(\phi')}{\gamma_s Z_{\max}\sin(\delta)\cos(\delta)} \tag{17}$$

Also, an extra condition (and TRIGRS equation) applies in addition to Eqs. 3 and 11–13:

$$F_f + F_w \geq 0 \tag{18}$$

It reflects the reduction of shear stress originating from the soil mass term F_f by some amount originating from the pore pressure term F_w. This latter entity cannot be larger than

F_f (e.g., [57]). Thus, $F_f + F_w$ is set to zero if it becomes smaller than zero, resulting in $F_s = F_c$. This way F_c acts as a lower bound constraint on F_s. The actual lower bound that we then implement is:

$$F_{\mathrm{s,lower}} = \max\left(F_{s,\,Z_{max},\,d\,=\,0}, F_{\mathrm{c}}\right) \tag{19}$$

When the F_s bounds in Eqs. 15 and 19 are on either side of the failure threshold of $F_s = 1$, we obtained the following lower bound on the valid d_{init} range from Eqs. 3 and 16 and $x = 1$:

$$d_{\mathrm{init},\,F_{\mathrm{s}}\,=\,1} = Z_{\max} - \frac{1}{\beta\gamma_{\mathrm{w}}\tan(\phi')}\left[c' - \left(1 - \frac{\tan(\phi')}{\tan(\delta)}\right)\gamma_{\mathrm{s}} Z_{\max}\sin(\delta)\cos(\delta)\right] \tag{20}$$

However, when $F_s = 1$ lies outside the F_s bounds range so that either $F_{\mathrm{s,lower}} \geq 1$ indicating a slope that is permanently stable or $F_{\mathrm{s,upper}} = F_c + F_s < 1$ indicating a slope that is always unstable, the corresponding value of the initial d_{init} does not matter. In such cases, our results become independent of the d_{init} range considered (we then arbitrarily consider a lower bound at $d_{\mathrm{init}} = 0$).

TRIGRS also necessitates definition of a steady pre-storm infiltration rate, I_{ZLT}, which determines the initial moisture content of our soil profile. For each site, we treat I_{ZLT} as an unknown parameter with uniform prior distribution bounded between zero and half of the maximum rainfall rate over the null and landslide failure event periods. This prior distribution thus takes into consideration site-specific variations in rainfall rates. Since the null events are selected from the three-month seasonal period before the landslide event, this method of assigning an uncalibrated distribution to I_{ZLT} is similar to the TRIGRS applicators' method of assigning it based on the mean rainfall over that seasonal period [4].

3.3.5 Hyperparameters—We assign a uniform distribution bounded between 0 and 1 for the hyperparameters, namely the False Omission Rate, R_{fo}, and the False Discovery Rate, R_{fd}, in Eqs. 5 and 6, respectively. These two statistics are related to the following contingency table metric of accuracy (A_c):

$$A_{\mathrm{c}} = \frac{n_{\mathrm{TN}} + n_{\mathrm{TP}}}{n_{\mathrm{FN}} + n_{\mathrm{TN}} + n_{\mathrm{FP}} + n_{\mathrm{TP}}} \tag{21}$$

where n_{FN}, n_{TN}, n_{FP} and n_{TP} are defined in Eqs. 5–6. The A_c value can range between 0 (poor performance) and 1 (perfect performance) so that higher values signify a better model performance. To explain A_c in terms of R_{fo} and R_{fd}, we specifically consider its complement (1-A_c) for which lower values are better:

$$1 - A_{\mathrm{c}} = \frac{n_{\mathrm{FN}} + n_{\mathrm{FP}}}{n_{\mathrm{FN}} + n_{\mathrm{TN}} + n_{\mathrm{FP}} + n_{\mathrm{TP}}} \tag{22}$$

Note that the sum of the numerators of the ratios in Eqs. 5–6 results in the numerator of Eq. 22. The same holds for the sum of the denominators of these two equations. The metric 1 −

A_c is thus simply a ratio of the sum of the numerators to the sum of the denominators of the individual R_{fo} and R_{fd} metrics. This metric must therefore be mathematically bounded by the minimum and maximum values of the two individual ratios:

$$\min(R_{fo}, R_{fd}) \leq 1 - A_c \leq \max(R_{fo}, R_{fd}) \tag{23}$$

3.3.6 Sampling uncalibrated correlated parameter sets using copulas—

Copulas enable a separate modeling of the marginals and the dependence (joint) structure of a multivariate distribution. They are multivariate-correlated probability distributions for which the marginal cumulative distribution functions (CDF$_s$) of the probabilities are uniform. This means that sampling from a multivariate distribution (like a standard normal one in our study) will provide vectors of CDF$_s$, where each vector is a grouping over TRIGRS parameters as an example, of the sampled values from the respective marginal CDF$_s$. For any TRIGRS parameter, these sampled marginal CDF values are invertible to provide the corresponding values of that parameter. Finally, we regroup these extracted parameter values over TRIGRS parameters to get the required vectors of sampled parameters.

This study assigned explicit correlations between only some parameter distributions, specifically for hydraulic parameters θ_r, θ_s, α, and K_s as mentioned in Section 3.3.1 further above. Note that the relevant elements of the specified correlation matrix of the copula need to be calculated from TRIGRS parameter values sampled from marginals that have the same distribution shape as the standard multivariate distribution used during the copula sampling phase. So, for this study using standard multivariate normal copula, the θ_r and θ_s do already come from normal distributions, and we log-transformed the values for α and K_s that come from lognormal distributions before calculating the relevant pairwise correlations.

We also have the information on applicable bounds on the TRIGRS parameter distributions (see Table 2). However these bounds can only be implemented through sampling of truncated distributions (for example, multivariate truncated standard normal copula will be relevant to our study). Sampling from such multivariate truncated distributions is considerably more difficult, and exact sampling is only feasible for truncation of the normal distribution to a polytope region [79, https://en.wikipedia.org/wiki/Truncated_normal_distribution]. For more general cases, a general methodology exists within a Gibbs sampling framework [80–81]. However, we did not implement any such truncation methodology in our study for our copula sampling. Theoretically, this can affect sampling from marginal distributions that have either bound close to the distribution mean instead of at the tails, since a large number of parameter values sampled from the unbounded distribution need to be discarded. However, not bounding the correlated parameters will have negligible impact on this study's results since θ_r, θ_s, α and K_s have bounds covering a wide valid-value range of 4σ per Table 2.

For the uncorrelated TRIGRS parameters, we were still able to sample from corresponding truncated marginal distributions by obtaining the bounded marginal CDF values from the unbounded marginal CDF values that were sampled from the copula (denoted by CDF$_s$

where subscript ‘s’ denotes sampled). This is done by first calculating CDF values at the lower and upper bounds ($\mathrm{CDF}_{l,u}$ and $\mathrm{CDF}_{u,u}$, respectively) in the unbounded distribution. Next, the CDF_{s} values having a 0–1 range are linearly mapped into the $\mathrm{CDF}_{l,u}$ — $\mathrm{CDF}_{u,u}$ range to obtain the corresponding bounded CDF values:

$$\mathrm{CDF}_b = (1 - \mathrm{CDF}_s)\mathrm{CDF}_{l,u} + \mathrm{CDF}_s\mathrm{CDF}_{u,u} \tag{24}$$

3.4 Stochastic TRIGRS calibration using the DREAM$_{(zs)}$ algorithm

The Stage IV rainfall estimates are typically considered to exhibit the highest CONUS-wide skill and finest spatial resolution compared to V7 and NLDAS-2; this product was considered as our ‘true’ forcing and used for calibration of the TRIGRS parameters. The resulting posterior parameter values were then forced separately with the V7 and NLDAS-2 rain time series to evaluate the sensitivity of simulated model output to each precipitation data product. We purposely do not calibrate TRIGRS with the V7 and NLDAS-2 rainfall products to avoid corruption of the model parameters by forcing data measurement errors.

The *a priori* distributions for our model calibration are the uncalibrated parameter distributions from Section 3.3. We infer the posterior distribution of the TRIGRS model parameters using the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm. Benchmark experiments (e.g., [58–62]) have demonstrated the ability of the DREAM algorithm to sample efficiently complex target distributions involving high-parameter dimensionality, multi-modality, and variably correlated, twisted and truncated parameters. In fact, practical experience suggests that DREAM often provides better solutions in high-dimensional parameter spaces than commonly used optimization algorithms [e.g., 87, 88]. In this paper, we implement the DREAM$_{(zs)}$ algorithm, a member of the DREAM family of algorithms, which uses parallel direction and snooker sampling from an archive of past states to evolve the different Markov chains to a stationary distribution [63]. This algorithm has the advantage of requiring only a few chains, independent of the number of model parameters. What is more, the use of diminishing adaptation (for convergence proof) allows for multi-core distributed evaluation of the candidate points in a manner which does not violate reversibility of the sampled chains.

The original DREAM$_{(zs)}$ is coded in MATLAB and necessitates the use a separate MATLAB license for each computational node in a distributed computing environment. This may be financially costly, and, hence, does not promote evaluation of the sampled chains of the large number of sites in our study assigned to multiple nodes. We therefore created a separate Python implementation of the DREAM$_{(zs)}$ algorithm. This Python code is a pared-down implementation of the MATLAB toolbox and monitors the convergence of the sampled chains using the univariate, $\hat{R}$ [64], and multivariate, $\hat{R}_d$ [65], diagnostics. We execute the Python code of the DREAM$_{(ZS)}$ algorithm with three different Markov chains using default values for the algorithmic variables. The initial size of the external archive was set to 280 (or 20 times the number of parameters in Table 2).

For each site, we judged the convergence of the DREAM$_{(zs)}$ algorithm using the $\hat{R}$ and $\hat{R}_{\mathrm{d}}$ diagnostics and verified whether the sampled posterior distribution contains at least one behavioral parameter vector which correctly simulates slope failure during a landslide period and slope stability during at least one of the null periods. Otherwise the TRIGRS simulation and associated parameter vector will be assigned a zero likelihood, commensurate with Eq. 8. For each of the 180 sites, we used a maximum total of 80,000 generations, which equates to a 80,000 × 3 = 240,000 TRIGRS model evaluations. For a handful of sites, however, this rather liberal computational budget was insufficient to find a sample with likelihood larger than zero.

To help locate behavioral solutions, we could have used a larger number of Markov chains with the DREAM$_{(zs)}$ algorithm. This is easy to do in practice, yet a more productive approach could have been to adapt Eq. 8 and differentiate among the likelihood values of the nonbehavioral solutions. This would introduce a gradient in the likelihood surface and help guide the DREAM$_{(zs)}$ algorithm to the behavioral solution set. We refer interested readers to the work of [89] who demonstrate that such more discriminatory likelihood function speeds up tremendously the efficiency of MCMC simulation within the context of Approximate Bayesian Computation. Nevertheless, we do not follow this alternative approach herein. Instead, we draw at random 90, 000 samples from the prior parameter distribution. For each site and sample from the prior distribution, we then execute TRIGRS and compute the likelihood function of Eq. 8. This rather simplistic approach led to at least one behavioral solution for three of the so-called outlier sites. These behavioral solutions are then combined with those derived from the DREAM$_{(zs)}$ algorithm and used as initial archive for a subsequent trial with this algorithm. This second time, we focus our attention only on those 168 sites with at least one behavioral solution (e.g., non-zero likelihood) and use a total of 40,000 generations to approximate the target distribution. Note, that for almost all sites, this second trial resulted in a similar approximation of the posterior distribution as the first run with DREAM$_{(zs)}$ algorithm. Samples after convergence are used to summarize the posterior distribution of the TRIGRS parameters.

3.5 Graphical analysis of results

At any site, we consider distributions of the binary simulated slope failure response over the null periods separately from those for the landslide failure period. The means of these distributions of binary response for the null periods and the landslide period, respectively, are ideally close to zero and to one. The landslide response distribution (i.e., during the landslide period) at a site simply is an ensemble in which each value is the individual binary response of a parameter set. However, since each site has multiple null events, we consider its null response distribution to be an ensemble in which each value corresponding to a parameter set is a simple average of the individual responses generated by that parameter set for the null periods. We consider the distribution characteristics of the mean and the spread of one standard deviation on either side of the mean for the relevant responses across the parameter sets. We further group the sites by SSURGO Musym soils, landslide event years and overall for display in Figs. 3 and 4, by elevation in Fig. 8, and by slope in Fig. 9 (to be detailed in Section 4).

For dynamic simulations of a spatially distributed model over a regional domain, computation of the performance metrics related to the contingency table (and in line with probability calculations by first principles) should involve counting over both the spatially and temporally discretized elements. However, the calculation of measures typically done in literature for such involves collapsing the time axis so that the measures reported involve counting only over the spatially discretized elements. The results of both approaches unavoidably depend on the assumed spatiotemporal discretization. Spatially, our study considers only the 'point' locations of slope failure and not the null locations, and we have considered null periods at the landslide locations instead. In our parameter uncertainty context where the ensemble of parameter sets (instead of a single set) is considered for calculating the relevant measures, the False Positive Rate (FPR; also called False Alarm Rate) is then the ratio of number of occurrences of simulated failure during observed null periods to that of all observed null periods. As it turns out, this is mathematically identical to the simulated mean of the null period response distribution mentioned above. Hence, FPR is mentioned on the y-axis labels of the first subplot in Figures 3, 4, 8, and 9. Similarly, the False Negative Rate (FNR; also called Miss Rate) is the ratio of number of occurrences of simulated stable slope during observed landslide periods to that of all observed landslide periods. The complement of FNR (i.e., True Positive Rate, or TPR=1-FNR) is mathematically identical to the mean of the landslide period response distribution mentioned above, and therefore 1-FNR is mentioned on the y-axis labels of the second subplot in Figures 3, 4, 8, and 9. An accuracy calculation (Eq. 21) would also be along the same lines: fraction of correctly simulated null and landslide periods to total number of null and landslide periods for the entire ensemble of parameter sets.

In Figs. 5–7 (to be further explained in Section 4), we plotted the parameter distributions in the same way as the binary response distributions mentioned above, but with the responses replaced by the parameter values or their normalized versions. Parameter normalizing for both the calibrated and uncalibrated values at each site to obtain minimum and maximum possible values of 0 and 1 is done by linearly scaling between the minimum and the maximum values of the uncalibrated distributions. Across all sites, any normalized parameter has the same shape of the uncalibrated distribution and the same multiple of the standard deviation defining the bounds of that distribution. Hence, the means and standard deviation ranges around the means of the normalized uncalibrated distributions for any parameter are the same across sites so that we simply denote them by horizontal black lines (solid and dashed respectively). Hyperparameters like R_{fo} and R_{fd} already range between 0 and 1 so that normalization does not change the original values (Fig. 5).

4. RESULTS

4.1 Rainfall pattern and modeled slope failure

Fig. 2 shows the rain time series averaged across sites for the TMPA V7, Stage IV, and NLDAS-2 data sources over the landslide and assumed null periods for landslides occurring in September 2004 and July 2011. For the 2004 events (Fig. 2a), the rain data from all sources show more rainfall volume during the landslide periods and less during the null periods. However, for the 2011 events (Fig. 2b), while Stage IV again provides this expected pattern of relative rainfall volumes between the landslide and null periods, the rain estimates from other sources of TMPA V7 and NLDAS-2 fail to do so and actually show a reverse pattern. The effect of such limitations in the rain accuracy of TMPA V7 and NLDAS-2 on the modeled responses are discussed below.

4.2 Uncalibrated TRIGRS performance and calibration relevance

We first evaluate the uncalibrated TRIGRS performance at any site by looking at the percentage of favorable parameter sets as defined in Section 3.4, out of a large fixed sample of 90,600 forming the uncalibrated distribution. That is, we consider only the parameter sets that simulate slope failure during the landslide period but simulate failure for only some and not all the null periods (to avoid parameter sets having a likelihood of 0 per Eq. 8). This percentage varies widely from 0% to almost 19%, where 22 sites show 0% and the rest mostly show less than 4%, while 10 sites do have high percentages of almost 19%. The mostly low percentages are near those seen in applications of hydrology and hydrologic hazard (e.g., <1% by [66] for semiarid flashflood modeling). The rest of the parameter sets (i.e., not the favorable parameter sets) simulate almost all sites as always being stable. The presence of 0% favorable parameter sets indicates the importance of a stochastic calibration procedure that allows directed sampling to converge to possibly tiny favorable regions in the parameter space.

The near-zero percentages reported above for most sites indicate the extremely intensive sampling required to delineate favorable portion/s in that space. This also explains why we used two subsequent DREAM$_{(zs)}$ trials for some of the sites as detailed in Section 3.4. The culprit is Eq. 8 which assigns a zero likelihood to all nonbehavioral solutions—no matter how far removed each solution is from the behavioral region. Such box-car type likelihood function introduces a zero gradient in large parts of the parameter space and makes it unnecessarily difficult for any search algorithm to locate the behavioral solution space. This is particularly true if the behavioral solution space occupies only a very small portion of the prior parameter distribution. Readers are referred to [89] for a remedy to Eq. 8. This approach would have made obsolete the two-step sampling approach used herein. Nevertheless, adding favorable initial parameter sets from a fixed intensive sampling at the start of the second DREAM$_{(zs)}$ run cycle enabled being in the non-zero likelihood regions of the parameter space where a gradient existed; hence, the direction of improvement in the likelihood function was followed. Note, that the likelihood function of Eq. 7 does not suffer the problems of Eq. 8, yet unavoidably results in a large cohort of parameter vectors that consistently simulated unstable slopes for all landslide and null periods (or at least for the landslide period and most of the null periods).

Next, regarding the characteristics of the uncalibrated response distributions, we also look at the groupings in Fig. 3 (refer Section 3.5 on how to interpret this Figure). Fig. 3 does not show any significant difference between the rain data sources, indicating that using alternate sources like V7 or NLDAS-2 negligibly changes the output response when using uncalibrated distributions in TRIGRS. Note that the only difference in the uncalibrated parameter set distributions between the sources is in the value of the pre-storm steady infiltration rate parameter (I_{ZLT}).

Fig. 3a shows the null period distributions (whose means coincide with the corresponding FPRs or False Positive Rates) for the three rain data sources, and for which the values should be ideally close to zero. Our simplistic quantitative analysis for interpreting whether slope failure is simulated by a response distribution involves checking whether the distribution means are above or below a threshold value of 0.5, since that is objectively midway between 0 and 1. The means are seen to be lower than 0.5 throughout and hence closer to zero than to one. For all rain sources during year 2004 and overall, the means or FPR values are equal to 0.07 and the standard deviational uncertainties lie in the 0.25–0.26 range. For year 2011, these numbers slightly change to means or FPR values in the 0.04–0.05 range and standard deviational uncertainties in the 0.2–0.21 range. These FPR mean values are significantly lower than the 0.22 or 22 % achieved by [82] for another physically based model, CHASM (Combined Hydrology and Slope Stability Model). They are also lower than the FPR achieved by GIS-based models (for example, 0.29 by [83] for the regional model).

Fig. 3b shows the distributions for the landslide period for which the values should ideally approach 1; however, all means (that coincide with the corresponding TPRs or True Positive Rates) are again lower than 0.5 and almost same as the distributions from the null periods of Fig. 3a. This further confirms the observation above that almost all parameter sets of the uncalibrated ensemble consistently simulate stable slopes for all periods. The response characteristics for all rain sources during landslide periods are similar to those for null periods: obviously unacceptable 0.04–0.1 values for the means or TPR values, and 0.2–0.23 for the standard deviational uncertainties.

If a landslide simulation model always simulates unconditionally stable slopes, the accuracy metric (A_c) can be misleading by always being high from reflecting the high percentage of stable slope elements among the total number of elements. For example, if our uncalibrated TRIGRS had always simulated slopes as unconditionally stable, only true negatives would have existed and not true positives: combining this information with our considered 5:1 ratio of number of null periods to landslide periods would have given A_c values of $^5/_6$ = 0.83. This is close to our obtained value of 0.79–0.80 (not shown). The consideration of such fixed ratios of numbers of periods in our study limits the utility of A_c for comparison against other studies, and for which we rely more on the FPR and TPR (or FNR) numbers below.

4.3 Performance of TRIGRS calibrated to Stage IV forcing

The calibrated response characteristics are shown in Fig. 4. This section focuses only on simulations forced with Stage IV data (green markers). Like Fig. 3a, Fig. 4a is for the null periods, again giving distribution means that are lower than 0.5. Calibration caused only small changes in the response means or FPRs for Stage IV forcing between Fig. 3a and Fig. 4a: from 0.07 to 0.06 for 2004, 0.05 to 0.27 for 2011, and 0.07 to 0.09 for overall. Compared to the uncalibrated value, the calibrated FPR value for 2011 is now in the ballpark of values attained by physically based and empirical GIS-based models (0.22 by [82] for CHASM, 0.29 by [84] for the regional model). The standard deviational uncertainty reduced from 0.26 to 0.17 for 2004 and from 0.25 to 0.2 overall, but increased from 0.21 to 0.29 for 2011. Here and in general, the overall response characteristics are closer to those for 2004 than for 2011 since only 22 out of the 168 sites failed in 2011.

Like Fig. 3b, Fig. 4b is for the landslide periods; however, calibration has now caused significant changes between them. The Stage IV response means or TPRs have now changed from 0.09 to 0.998 for 2004, 0.06 to 0.83 for 2011 and 0.09 to 0.98 for overall, putting them above the 0.5 threshold and close to 1 as desired. Compared to the low uncalibrated TPR values reported in Section 4.2, these calibrated TPR values are now better than the numbers obtained by other physically based models (e.g., 0.68 by [82] for CHASM, 0.71 by [84] for SINMAP). The standard deviational uncertainty reduced from 0.29 to 0.04 for 2004 and from 0.28 to 0.15 overall, but increased from 0.23 to 0.37 for 2011. The accuracy values (A_c) are now high due to calibration for 2004 (0.95) and overall (0.92), but have actually slightly decreased to from 0.8 to 0.75 for 2011 (not shown).

For Stage IV data during 2011 (when compared against 2004 and overall), the reporting above of higher response mean values during the null periods and the consistent increase in standard deviational uncertainty during both landslide and null periods is an indication of possible difficulty in capturing the true rain estimates during 2011. Also, we note from Fig. 4b that the calibration procedure was unsuccessful for the '*Ss*' soil (Spivey-Santeetlah-Nowhere complex, very stony, and Spivey-Santeetlah complex, very stony), since its Stage IV mean is below the 0.5 threshold. It is also somewhat close to but above 0.5 for the '*Ow*' soil (Oconaluftee channery loam). The standard deviational uncertainties for these two soils are also high.

To recap, the results from this and Section 4.3 answer one of our research questions about the relevance of calibration by showing that calibration provides significantly improved modeled response over the uncalibrated model. In fact, without calibration, it can be difficult to have any useful simulations of slope stability in an uncertainty framework. The rest of the figures and analyses in this study consider only the calibrated simulations.

4.4 Sensitivity of calibrated simulations to rain data source

This section focuses only on Fig. 4. Fig. 4a for the null periods shows only minute changes for alternate rain sources in the response means or FPRs from the Stage IV values during 2004 and overall (0.11 for V7 and 0.07–0.1 for NLDAS-2 versus 0.06–0.09 for Stage IV). However, the V7 response mean for 2011 at 0.08 shows a noticeable difference from Stage IV and NLDAS-2 values at 0.27. Similarly, in the response standard deviational uncertainties, there are only minute changes for alternate rain sources from the Stage IV values during 2004 and overall (0.23 for V7 and 0.2–0.22 for NLDAS-2 versus 0.17–0.2 for Stage IV). However, the V7 standard deviational uncertainty for 2011 at 0.2 shows a noticeable difference from Stage IV and NLDAS-2 values at 0.29.

Fig. 4b for the landslide periods shows that Stage IV has the best performance value for the Stage IV response means or TPRs (0.998 for 2004, 0.83 for 2011 and 0.98 for overall) since this is the only rain source data that was actually used for calibration. For year 2004 and overall, the next best performance is that for NLDAS-2 (0.88 for 2004, 0.77 for overall) and then V7 (0.8 for 2004, 0.7 for overall): V7 and NLDAS-2 values are much above the 0.5 threshold and so acceptably resolve the peak rainfall during the landslide period of the 2004 storms. However, for year 2011, the performances of both NLDAS-2 and V7 are unacceptable (0.02 for NLDAS-2, 0.06 for V7): these data sources failed to characterize the rainfall associated with the landslide event period during 2011. Comparisons of the standard deviational uncertainties for year 2004 and overall are as expected: Stage IV has significantly lower standard deviational uncertainty (0.04 for 2004, 0.15 for overall) than V7 (0.4 for 2004, 0.46 for overall) and NLDAS-2 (0.32 for 2004, 0.42 for overall). However, for 2011, the standard deviational uncertainty for Stage IV has jumped to a much higher value at 0.37 than at 0.24 for V7 and 0.15 for NLDAS-2. For V7 and NLDAS-2 during the landslide periods of 2011, the simultaneous consideration of means close to zero and the low standard deviations shows that these rain data sources provide very low chance of slope failure at these sites.

Both V7 and NLDAS-2 rain data sources could not provide satisfactory performance (i.e., calibrated means below 0.5) for four soils including the *Ss* and *Ow* soils mentioned in Section 4.4, the other two being '*Hc*' (Heintooga-Chiltoskie complex, stony) and '*Oc*' (Oconaluftee channery loam). Coincidently, only these four soils are present at the sites that failed during the 2011 rain event. Additionally, the 2011 rains correspond to all the *Ss*, *Ow* and *Hc* soil sites, and to 15 out of 16 *Oc* soil sites (the sole *Oc* soil site in 2004 does provide a near-perfect calibrated performance for all rain sources). This shows the difficulty faced by any slope stability model in simulating failures when the rain data quality is inadequate. Finally, while there are many soils where V7 means seems to perform noticeably worse than those of Stage IV and NLDAS-2 (some examples are *Ac*, *Bw, Cn, Ct, Uk*), the V7 simulations are not useful (i.e., mean is below 0.5) for two soils (*Ew:* Evard-Cowee complex, stony; and *To:* Toecane-Tusquitee complex, bouldery). This strongly indicates the quality of the V7 data being worse than that of NLDAS-2 for the 2011 rains.

We can now provide the answer to this study's overarching question of change in modeled responses with rain source: the response means for V7 and NLDAS-2 do not show any appreciable degradations against Stage IV during periods of good rain data quality (i.e., when relative rain volumes between null and landside periods are as expected), so that V7 and NLDAS-2 can provide successful simulations using the calibrated model. However, the responses completely degrade during periods of poor rain data quality for V7 and NLDAS-2 (i.e., when relative rain volumes between null and landside periods follow the opposite pattern) so that slope failures cannot be effectively predicted even with the calibrated model.

4.5 Model parameter distributions

We first check the results of the calibration procedure by using Eq. 23 relating R_{fo}, R_{fd} and A_{c}. Fig. 5 shows the means of calibrated distributions (refer Section 3.5 on how to interpret this figure) of hyperparameters R_{fo} and R_{fd} to mostly be below 0.5 (i.e., except for the '*So*' or Soco-Stecoah complex soil, and almost 0.5 for the '*To*' or boulder Toecane-Tusquitee complex soil and the *Ow* soil). This indicates that the R_{fo} and R_{fd} values are now mostly below 0.5: R_{fo} means are now at 0.36 for 2004, 0.42 for 2011 and 0.37 for overall, while f means have now reached 0.37 for 2004, 0.44 for 2011 and 0.38 for overall. Hence, it follows that $1 - A_{\mathrm{c}}$ is also mostly below 0.5 and closer to zero (as desired) than to 1. In other words, the calibration procedure has resulted in accuracy (A_{c}) values closer to 1.

Next, we look at the initial conditions. Fig. 6a shows the distributions for the steady pre-storm infiltration rate (I_{ZLT}), which is the only parameter that differs in value between uncalibrated simulations forced with Stage IV, V7 and NLDAS-2, respectively. For uncalibrated distributions, NLDAS-2 (red) means and standard deviation ranges are mostly lower than those for Stage IV (green) and V7 (blue). Similarly, Stage IV calibrated distributions (black) have also shifted towards lower values when compared to Stage IV uncalibrated distributions. In other words, our assumption of upper-bounding the I_{ZLT} distribution at half of the maximum rainfall rate is an overestimation (refer Section 3.3.4 on initial conditions). Fig. 6b shows that the calibrated distributions for the normalized initial depth to the water table (d_{init}) have substantially lower means than those for the Stage IV uncalibrated distributions (there is reduction by 0.28 for 2004, 0.37 for 2011, and 0.29 for overall). Calibration has also reduced the standard deviational uncertainty in the normalized d_{init}, by 0.07 for 2004, 0.14 for 2011 and 0.08 for overall. This indicates that the soil columns clearly are substantially wetter than in our assumed uncalibrated d_{init} distributions.

Depending on the parameter considered, the Stage IV uncalibrated distribution means can be underbiased or overbiased when compared to the calibrated means. For example, Fig. 7a shows calibration correcting the example geotechnical parameter of the effective cohesion (c') towards lower means (the normalized values have reduced by 0.15 for 2004, 0.08 for 2011 and 0.14 for overall). Subjectively considering a minimum reduction in normalized means by more than 0.05 for overall, the other TRIGRS parameters that are corrected towards lower means are the soil friction angle for effective stress (ϕ': reductions by 0.05 for 2004 and overall, and 0.02 for 2011), and the specific storage (S_{s}: reductions by 0.05 for 2004, 0.08 for 2011 and 0.06 for overall). Similarly considering a minimum increase in normalized means by more than 0.05 for overall, the TRIGRS parameters corrected towards higher means (not shown) are the geomorphologic parameters of the depth to bedrock (Z_{max}: 0.11 for 2004 and overall, and 0.06 for 2011) and the slope (δ: 0.09 for 2004, 0.05 for 2011 and 0.08 for overall). The remaining parameters do not show noticeable change due to calibration by having deviation in the normalized means that is less than our subjective threshold of 0.05 for overall. Fig. 7b shows calibration leaving the distribution characteristics for such an example hydraulic parameter like the saturated soil moisture content (θ_{s}) almost unchanged. Other parameter means that were relatively unchanged are the hydraulic parameters like the residual soil moisture content (θ_{r}), the saturated hydraulic conductivity (K_{s}) and the inverse of the capillary rise (α), and the geotechnical parameter of the unit weight of soil (γ_{s}). These uncalibrated distributions for these unchanged parameters have been derived from SSURGO and ROSETTA for the hydraulic parameters, and NAVFAC DM 7.01 manual [40] for γ_{s}, indicating the high accuracy of these sources and low sensitivity of F_{s} to γ_{s}.

For standard deviational uncertainty of the normalized distributions, we similarly consider a subjective threshold change of 0.025 due to calibration so that the corresponding threshold change in spread is 0.05 (the spread spanning either side of the mean is twice the standard deviation). Considering standard deviational uncertainty of the hydraulic parameters, we find that the change is below this threshold, as in the changes reported above for the corresponding normalized means (but now S_{s} also shows negligible change in the standard deviational uncertainty). Parameters that exhibit changes more than the standard deviation threshold are the geotechnical parameters of c' (increase by 0.08 for 2004 and overall, and 0.06 for 2011), ϕ' (increase by 0.03 for 2004 and overall, and 0.02 for 2011) and γ_s (a consistent increase by 0.3 for 2004, 2011 and overall), and the geomorphologic parameter of $Z_{\max}$ (increase by 0.04 for 2004 and 0.03 for overall but decrease by 0.02 for 2011). The mostly slight increase of standard deviation for these parameters shows that calibration has corrected their uncalibrated distributions towards flatter ones.

In general, the changes in the normalized means and standard deviations reported above now enable us to answer our research question about changes in parameter distributions due to calibration: calibration is typically not required for the hydraulic parameters derived from SSURGO and ROSETTA. Geotechnical and geomorphologic parameters requiring calibration to correct both the mean bias and the standard deviation are c', ϕ' and $Z_{\max}$. Additionally, geomorphologic δ requires some correction to the mean and geotechnical γ_s requires some correction to the standard deviation.

4.6 Modeled response for ranges of elevation and slope

To determine the relative sensitivity of model results to differences in elevation or slope, we group the characteristics of the calibrated model response distributions by elevation or slope ranges. Instead of grouping the elevations for all sites into equal-interval bins as is typically done for obtaining the ranges, we group the sites into equal-frequency bins [67] so that the range varies among these bins. Such adaptive binning avoids relative undersampling-based erroneous statistics for any bin, essentially becoming optimized estimators of information for constructing histograms [68–69].

Considering grouping of sites by elevation ranges, Fig. 8a for null periods shows the response means or FPRs being desirably below the 0.5 threshold. However, Fig. 8b for landslide periods (where desirable performance is above the 0.5 threshold) shows a clear trend for V7 and NLDAS-2 rains of modeled response values moving towards zero at the highest elevation ranges. In fact, these rain sources mostly give response mean values close to zero for the highest elevation bin having a range of 5062–5828 ft, meaning that this range mostly contains simulation outputs where slopes are stable for both the null and landslide periods. Satellite rainfall estimates can have difficulty resolving orographically enhanced rainfall in higher elevations due to the warm rain processes typical in these environments, which essentially can cause the satellite to underestimate the rainfall associated with the brightness temperatures of these types of storms relative to comparable precipitating systems in flat terrain [70–72]. The satellite estimates can be further biased due to the limited gauge network within higher elevations and complex topography that is used to calibrate the V7 research product. Finally, the coarser spatial resolution of the V7 and NLDAS-2 products may impact the accurate characterization of local heterogeneities (e.g. higher localized rainfall values) in complex terrain.

Delving into the effect of elevation, the highest elevation range of 5062–5828 ft covers 17 sites. Out of these, a dominant number of 13 sites (or 76 % of the sites) have recorded slope failure occurring during the 2011 period of poor rain data quality for V7 and NLDAS-2 (refer to Section 4.1). The next highest elevation range of 4600–5062 ft has a more balanced proportion of 8 out of a total of 18 sites (or 44% of the sites) where slope failures were recorded during 2011. This corresponds to the modeled response near 0.5 for V7 and NLDAS-2 in this elevation range showing less domination by the 2011 rains (Fig. 8b).

Similarly, the third-highest elevation range of 4230–4600 ft containing the remaining 4 of the total of 22 sites with 2011 rains has a low proportion of the total 16 sites (4/16 = 25%), and this is reflected in the response means for V7 and NLDAS-2 being even closer to 1 than higher elevation ranges. This strongly suggests the relationship between higher elevations and poorer rain data quality for V7 and NLDAS-2 during the 2011 period. Therefore, we can now answer our research question on the connection between elevation and modeled slope failure: higher elevations can impact rain data quality for V7 and NLDAS-2, which can significantly decrease the success of corresponding landslide simulations within these regions.

Fig. 9 showing the responses for different slope bins reflects no clear trend in the response means or TPRs (i.e., 1-FNRs) per Fig. 9b. The 34–35° bin where both V7 and NLDAS-2 are below the 0.5 threshold has a total of nine sites. Five of these sites (or 56 % of the sites) have rains that occurred in 2011. The 26–29° and 29–31° bins each have the V7 rain response means just below 0.5 even though the corresponding NLDAS-2 values are above 0.5. These two bins each have a total of 13 sites, out of which 5 sites (or 31% of the sites) showed failure during the 2011 rains. The 33–34° bin where both V7 and NLDAS-2 show response means close to but above 0.5 has 5 sites that showed failure in 2011 out of a total of 16 (or 31% of the sites). Remaining sites showing failure during rains in 2011 are situated in the 31–33° bin (3 sites out of total of 17, or 18% of the sites) and the 39–41° bin (1 site out of a total of 15, or 7% of the sites).

5. DISCUSSION

The research question addressed in this paper is the extent to which coarser resolution rain products like V7 or NLDAS-2 can be used in lieu of finer resolution data like gauge or Stage IV within a deterministic slope stability model. The results of this study demonstrate that for hillslope scale evaluation of slope stability via deterministic simulation (single parameter vector), only a finer resolution product like Stage IV may consistently and accurately resolve slope failures. However, the results also show that for particular cases, such as the 2004 storms that do not have the loss in data quality associated with factors like higher elevations, the V7 and NLDAS-2 rainfall estimates show promise in providing comparable results to the ‘truth’ used to resolve slope failure. The 2011 storm occurred at the higher elevations where V7 and NLDAS-2 rainfall data quality was greatly compromised due to orographically enhanced rainfall for warm rain processes. This type of rainfall is particularly difficult to determine accurately due to underestimation of brightness temperatures, the limited gauge network in complex topography and the coarser spatial resolution that does not characterize very well local heterogeneities in complex higher-elevation terrain as finer resolution products do.

In this study, we investigate this and related research questions using numerical simulation with an ensemble of parameter vectors derived from Bayesian inference using as our reference the Stage IV rainfall product with relatively fine spatial resolution. We find that the results are dependent on the quality of the rainfall data which tends to degrade at higher elevations. There is hardly any sensitivity to rainfall product during periods with adequate V7 and NLDAS-2 precipitation estimates. For those periods, these two rainfall products are of sufficient quality to warrant landslide prediction. However, the accuracy with which V7 and NLDSA-2 simulate hillslope stability deteriorates rapidly during periods where the peak rainfall values associated with a landslide-triggering storm are entirely or predominantly missed by the two products, which this study finds as being more prevalent at higher elevations.

While results suggest that overall V7 and NLDAS-2 may provide comparable performance to Stage IV if implemented at lower elevations within our study region, an improvement in two areas can make modeled response reliably insensitive to rainfall product: (i) improvement in quality of data from coarse resolution rain data sources towards acceptably low error at higher resolutions, and (ii) many more periods of recorded landslide observations (and over more regions) to provide further credence to characterize dependence on elevation of the trend in rainfall error and consequent modeled response. Also, the requirement of TRIGRS calibration in this study points to the importance of improvement in uncalibrated parameter distributions using in situ geomorphologic measurements of the depth to bedrock, and in research into the development of pedotransfer functions for geotechnical parameters.

To the best of our knowledge, our study is the first to simultaneously consider spatial variation of uncertainty in hydraulic, geotechnical and geomorphologic parameters over multiple locations in a region. The application of a physically based TRIGRS model in this study in an unsaturated and finite-depth mode is in contrast with most earlier studies where only saturated or quasi-saturated conditions were simulated in the soil profile following the historical legacy of slope stability modeling (e.g., [14]).

Our study also introduced a likelihood function that can be used to evaluate binary model output over a spatiotemporal domain. This function is continuous and combines information from landslide and null periods to improve parameter estimation and spatiotemporal characterization of slope stability. This likelihood function is easy to implement and use and can be applied to a suite of other simulation models, be it physically based or empirical, point-based or spatially distributed, or event-based or continuous. The selection of null and landslide periods is easily automatable for continuous simulation models and for event-based models that have the flexibility to start in the dry days leading up to a storm event or at the onset of rainfall. The proposed likelihood function may also be of use to other applications with categorical data, for example, in hydrologic hazards like semiarid flashflood modeling where events can be scarce and the discharge volume above some threshold is of interest. The likelihood function can be refined per our suggestions to further improve search efficiency and reduce as much as possible the required computational budget for parameter estimation. This would simplify the calibration of CPU-intensive, distributed, hillslope stability models, which is of particular importance in large-scale applications.

Note that the sensitivity tackled in this paper is the sensitivity to nominal discrete categories (i.e., not having any progression or rank between them) like rain data sources, and this type of sensitivity approach typically involves comparing the simulations between the categories. This is different from the typical sensitivity analyses done in literature on cardinal numbers (i.e., indicating quantity) that can be continuous or discrete in nature. Ensemble-based sensitivity analyses of this latter type of sensitivity analyses are usually fixed-sampling approaches like regionalized and globalized sensitivity approaches that sample from distributions of the cardinal variables to give some information of contributions of the different factor uncertainties to the uncertainty in the continuous output (e.g., [85, 66]). For discrete output observations as in our study that have not been converted to continuous values and so essentially constitute a mapping problem, the relevant type of approaches include the example Classification and Regression Trees (CART) procedure (e.g., [11]). This CART procedure considers only the binary response to optimize towards a high percentage of failed or stable slopes, and leads to determination of a parsimonious classification tree and its combinations of critical thresholds from the uncalibrated distributions of the model parameters.

We hope that our use of a large number of sites and stochastic simulations in this study helps to advance progress in the field of dynamic modeling of landslides. Additional studies over other regions and where further consideration of uncertainties in the rain forcing and even in model structure are needed to advance potential applications of satellite estimates within these types of model studies.

ACKNOWLEDGEMENTS:

This work was partially funded by the following NASA grants: NNH11ZDA001N-NIP (PI: Dr. Dalia Kirschbaum), NNH15ZDA001N-PMM (PI: Dr. Dalia Kirschbaum) and NNX12AO19G (PI: Prof. Ning Lu). We acknowledge David Giordano (NC OneMap Database Administrator) for providing us the 20-ft DEM for the entire state of North Carolina on August 4, 2015, Marcel Schaap (University of Arizona) for providing the ROSETTA Python code, and Thomas Stanley for help with linking the SSURGO tables in Microsoft Access. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES:

1. Croizer MJ, 1986: Landslides: Causes, Consequences and Environment. Croom Helm Ltd., Kent, UK 252 pp.
2. Sidle RC, and Ochiai H, 2006: Landslides: Processes, Prediction, and Land Use Water Res Monogr Ser 18, AGU, Washington D. C., 312 pp.
3. Salciarini D, Godt JW, Savage WZ, Baum RL, and Conversini P, 2008: Modeling landslide recurrence in Seattle, Washington, USA. Eng. Geol, 102, 227–237, doi:10.1016/j.enggeo.2008.03.013.
4. Baum RL, Godt JW, and Savage WZ (2010), Estimating the timing and location of shallow rainfall - induced landslides using a model for transient, unsaturated infiltration, J. Geophys. Res, 115, F03013, doi:10.1029/2009JF001321.
5. Apip K. Takara Y. Yamashiki K. Sassa AB Ibrahim, and Fukuoka H, 2010: A distributed hydrological-geotechnical model using satellite-derived rainfall estimates for shallow landslide prediction system at a catchment scale. Landslides, 7, 237–258, doi:10.1007/s10346-010-0214-z.
6. NCAR/EOL/NSF, 2016: NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Precipitation Data provided by NCAR/EOL under sponsorship of NSF. Online at http://data.eol.ucar.edu/codiac/dss/id=21.093
7. Huffman GJ, and Coauthors, 2007: The TRMM Multi satellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. J. Hydrometeorol, 8, 38–55, doi:10.1175/JHM560.1.
8. Huffman GJ, and Bolvin DT, 2012: Real-Time TRMM Multi-Satellite Precipitation Analysis Data Set Documentation. Online at ftp://trmmopen.gsfc.nasa.gov/pub/merged/V7Documents/3B4XRT_doc_V7.pdf

9. 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. J. Geophys. Res 117, D03109.
10. Xia Y, Mitchell K, Ek M, Cosgrove B, Sheffield J, Luo L, Alonge C, Wei H, Meng J, Livneh B, Duan Q, Lohmann D, 2012 Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. J. Geophys. Res 117, D03110.
11. Almeida S, Holcombe EA, Pianosi F, and Wagener T: Dealing with deep uncertainties in landslide modelling for disaster risk reduction under climate change, Nat. Hazards Earth Syst. Sci, 17, 225–241, doi:10.5194/nhess-17-225-2017, 2017.
12. Gomes GJC, Vrugt JA, Vargas EA Jr., Camargo JT, Velloso RQ, and Th. Van Genuchten M. (2017), The role of uncertainty in bedrock depth and hydraulic properties on the stability of a variably-saturated slope, Computers and Geotechnics, 88, 222–241, 10.1016/j.compgeo.2017.03.016.
13. Hamm NAS, Hall JW, and Anderson MG (2006), Variance-based sensitivity analysis of the probability of hydrologically induced slope stability, Computers & Geosciences, 32, 803–817, doi:10.1016/j.cageo.2005.10.007.
14. Frattini P, Crosta G, Sosio R. 2009 Approaches for defining thresholds and return periods for rainfall-triggered shallow landslides. Hydrological Processes 23: 1444–1460. DOI:10.1002/hyp.7269.
15. Arnone E, Dialynas YG, Noto LV, and Bras RL: Accounting for soils parameter uncertainty in a physically based and distributed approach for rainfall-triggered landslides, HydrolProcess, 30, 927–944, doi:10.1002/hyp.10609, 2016.
16. Gupta HV, Perrin C, Blöschl G, Montanari A, Kumar R, Clark M, and Andréassian V, 2014: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci, 18, 463–477, doi:10.5194/hess-18-463-2014.
17. Wooten RM, Gillon KA, Witt AC, Latham RS, Douglas TJ, Bauer JB, Fuemmeler SJ, Lee LG (2008), Geologic, geomorphic, and meteorological aspects of debris flows triggered by Hurricanes Frances and Ivan during September 2004 in the Southern Appalachian Mountains of Macon County, North Carolina (southeastern USA), Landslides, 5:31–44, DOI 10.1007/s10346-007-0109-9.
18. Soil Survey Staff (2015), Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database. Available online at http://sdmdataaccess.nrcs.usda.gov/ Accessed [09/2015]
19. Lin Y. 2011 GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data, Version 1.0. UCAR/NCAR - Earth Observing Laboratory. http://data.eol.ucar.edu/dataset/21.093. Accessed 26 Sep 2016.
20. Huffman GJ, Adler RF, Bolvin DT, Nelkin EJ, 2010 The TRMM Multi-satellite Precipitation Analysis (TMPA) In: Hossain F, Gebremichael M. (Eds.), Satellite Rainfall Applications for Surface Hydrology. Springer Verlag, pp. 3–22.
21. Chen M, Shi W, Xie P, Silva VBS, Kousky VE, Wayne Higgins R, Janowiak JE, 2008 Assessing objective techniques for gauge-based analyses of global daily precipitation. J. Geophys. Res. Atmos 113.
22. Xie P, Yatagai A, Chen M, Hayasaka T, Fukushima Y, Liu C, Yang S, 2007 A Gauge-Based Analysis of Daily Precipitation over East Asia. J. Hydrometeorol 8, 607.
23. Daly C, Neilson RP, Phillips DL, 1994 A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. J. Appl. Meteorol 33, 140–158.
24. Joyce RJ, Janowiak JE, Arkin PA, Xie P, 2004 CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrometeorol 5, 487–503.

25. Higgins RW, Janowiak JE and Yao Y, 1996: A gridded hourly precipitation data base for the United States (1963–1993). NCEP/Climate Prediction Center Atlas No. 1. Online at http://www.cpc.ncep.noaa.gov/research_papers/ncep_cpc_atlas/1/index.html
26. Mesinger F, and Coauthors, 2006: North American Regional Reanalysis. Bull. Amer. Meteor. Soc, 87 (3), 343–360. doi:10.1175/BAMS-87-3-343
27. Alvioli M, and Baum RL (2016), Parallelization of the TRIGRS model for rainfall-induced landslides using the message passing interface, Environmental Modelling & Software, 81: 122–135, 10.1016/j.envsoft.2016.04.002
28. Srivastava R, and Yeh T-CJ, 1991: Analytical solutions for one-dimensional, transient infiltration toward the water table in homogeneous and layered soils, Water Resour. Res, 27, 753–762.
29. Taylor DW, 1948: Fundamentals of Soil Mechanics. Wiley, New York (ISBN-13: 978–1258768928). 700pp.
30. Lu N, and Godt J. (2008), Infinite slope stability under steady unsaturated seepage conditions, Water Resources Research, 44: W11404, doi:10.1029/2008WR006976.
31. Schaap MG, Leij FJ, and Th M. van Genuchten, 1998, “Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity”, Soil Science Society of America Journal, vol. 62, pp. 847–855.
32. Schaap MG, Leij FJ, and Th M. van Genuchten, 2001, “Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions”, Journal of Hydrology, vol. 251, pp. 163–176.
33. Melchiorre C, Frattini P. 2012 Modelling probability of rainfall-induced shallow landslides in a changing climate, Otta, Central Norway. Climatic Change 113: 413–436. DOI:10.1007/s10584-011-0325-0.
34. Benson CH 1993 Probability distributions for hydraulic conductivity of compacted soil liners. Journal of Geotechnical and Geoenvironmental Engineering 119(3): 471–186.
35. Iverson RM, and Major JJ (1987), Rainfall, ground-water flow, and seasonal movement at Minor Creek landslide, Northwestern California—Physical interpretation of empirical relations, Geol. Soc. Am. Bull, 99, 579–594.
36. Baum RL, and Reid ME (1995), Geology, hydrology, and mechanics of a slow-moving, clay-rich landslide, Honolulu, Hawaii, Rev. Eng. Geol, vol. X, pp. 79–105 Geol. Soc. Am., Boulder, Colo.
37. Lumb P. 1974 Applications of statistics in soil mechanics In: Soil Mechanics. New Horizons, JK L (ed). American Elsevier: London.
38. Scharnagl B, Vrugt JA, Vereecken H, Herbst M. Inverse modelling of in situ soil water dynamics: investigating the effect of different prior distributions of the soil hydraulic parameters. Hydrol Earth Syst Sci 2011;15(10):3043–59. 10.5194/hess-15-3043-2011
39. Davis ROE, Bennett HH (1927) “Grouping of soils on the basis of mechanical analysis.” United States Department of Agriculture Departmental Circulation No. 419.
40. Warrington DC (1986), Soil Mechanics: NAVFAC Design Manual 7.01, United States Naval Facilities Engineering Command, 380 pp.
41. Langejan A. 1965 Some Aspects of the Safety Factor in Soil Mechanics Considered as a Problem of Probability. Sixt Int. Conf. Soil Mechanics and Foundation Eng: Montreal, Quebec, 500–502.
42. Tobutt DC. 1982. Monte Carlo simulation methods for slope stability. Computers & Geosciences 8: 199–208. DOI:10.1016/0098-3004(82)90021-8,1982.
43. Rackwitz R. 2000 Reviewing probabilistic soils modelling. Computers and Geotechnics 26: 199–223. DOI:10.1016/S0266-352X(99)00039-7.
44. Simoni S, Zanotti F, Bertoldi G, Rigon R. 2008 Modelling the probability of occurrence of shallow landslides and channelized debris flows using GEOtop-FS. Hydrological Processes 22: 532–545. DOI:10.1002/hyp.6886.
45. Dettinger M, Wilson J. 1961 First order analysis of uncertainty in numerical models of groundwater flow part: 1. Mathematical development. Water Resources Research 17: 149–161.
46. Matsuo M, Kuroda K. 1974 Probabilistic approach to design of embankments. Soils and Foundations 14: 1–17.

47. Christian J, Ladd C, Baecher G. 1994. Reliability applied to slope stability analysis. Journal of Geotechnical Engineering 120: 2180–2207. DOI:10.1061/(ASCE)0733-9410(1994)120:12(2180).

48. Malkawi AI, Hassan WF, Abdulla FA. 2000 Uncertainty and reliability analysis applied to slope stability. Structural Safety 22: 161–187.

49. Abbaszadeh M, Shahriar K, Sharifzadeh M, Heydari M. 2011 Uncertainty and reliability analysis applied to slope stability: a case study from Sungun copper mine. Geotechnical and Geological Engineering 29: 581–596. DOI:10.1007/s10706-011-9405-1.

50. Lumb P. 1970 Safety factors and the probability distribution of soil strength. Canadian Geotechnology Journal 7: 225–242.

51. DeRose RC, Trustrum NA, and Blaschke PM (1991), Geomorphic change implied by regolith-slope relationships on steepland hillslopes, Taranaki, New Zealand, Catena, 18(5), 489–514, doi:10.1016/0341-8162(91)90051-X.

52. DeRose RC, 1996: Relationships between slope morphology, regolith depth, and the incidence of shallow landslides in eastern Taranaki Hill Country, Z. Geomorphol. Supp, 105, 49–60.

53. Boer M, Barrio GD, and Puigdef_abres J. (1996), Mapping soil depth classes in dry Mediterranean areas using terrain attributes derived from a digital elevation model, Geoderma, 72(1–2), 99–118, doi :10.1016/0016-7061(96)00024-9.

54. Ziadat F. (2010), Prediction of soil depth from digital terrain data by integrating statistical and visual approaches, Pedosphere, 20(3), 361–367, doi:10.1016/S1002-0160(10)60025-2.

55. Wilford J, and Thomas M. (2013), Predicting regolith thickness in the complex weathering setting of the central Mt Lofty Ranges, South Australia, Geoderma, 206, 1–13, doi:10.1016/j.geoderma.2013.04.002.

56. Yang Q, Zhang F, Jiang Z, Li W, Zhang J, Zeng F, and Li H. (2014) Relationship between soil depth and terrain attributes in karst region in Southwest China, J. Soils Sediment, 14(9), 1568–1576, doi:10.1007/s11368-014-0904-6.

57. Duncan JM, Wright SG, and Brandon TL 2014 Soil Strength and Slope Stability. John Wiley & Sons 336 pp.

58. Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, and Robinson BA, “Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation,” Water Resources Research, vol. 44, W00B09, doi:10.1029/2007WR006720, 2008

59. Vrugt JA, ter Braak CJF, Diks CGH, Higdon D, Robinson BA, and Hyman JM, “Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling,” International Journal of Nonlinear Sciences and Numerical Simulation, vol. 10, no. 3, pp. 273–290, 2009.

60. Laloy E, Rogiers B, Vrugt JA, Jacques D, and Mallants D, “Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion,” Water Resources Research, vol. 49 (5), pp. 2664–2682, doi :10.1002/wrcr.20226, 2013.

61. Linde N, and Vrugt JA, “Distributed soil moisture from crosshole ground penetrating radar travel times using stochastic inversion,” Vadose Zone Journal, vol. 12 (1), doi:10.2136/vzj2012.0101, 2013.

62. Lochbuhler T, Breen SJ, Detwiler RL, Vrugt JA, and Linde N, “Probabilistic electrical resistivity tomography for a CO2 sequestration analog,” Journal of Applied Geophysics, vol. 107, pp. 80–92, doi:10.1016/j.jappgeo.2014.05.013, 2014.

63. ter Braak CJF, and Vrugt JA (2008). Differential Evolution Markov Chain with snooker updater and fewer chains. Statistics and Computing. 10.1007/s11222-008-9104-9

64. Gelman AG, and Rubin DB, “Inference from iterative simulation using multiple sequences,” Statistical Sciences, vol. 7, pp. 457–472, 1992.

65. Brooks SP, and Gelman A, “General methods for monitoring convergence of iterative simulations,” Journal of Computational and Graphical Statistics, vol. 7, pp. 434–455, 1998.

66. Yatheendradas S, Wagener T, Gupta H, Unkrich C, Goodrich D, Schaffner M, and Stewart A, 2008: Understanding uncertainty in distributed flash flood forecasting for semiarid regions, Water Resour. Res, 44, W05S19, doi:10.1029/2007WR005940.

67. Kraskov A, Sto'gbauer H, Grassberger P. (2004), Estimating mutual information, Physical Review E 69: 066138.

68. Darbellay GA and Vajda I, Estimation of the Information by an Adaptive Partitioning of the Observation Space, IEEE Trans. Inf. Theory 45, 1315 (1999).

69. Fraser AM and Swinney HL (1986), Independent coordinates for strange attractors from mutual information, Phys. Rev. A, 33, 1134.

70. Barros AP, and Lettenmaier DP, 1994: Dynamic modeling of orographically induced precipitation. Rev. Geophys, 32, 265–284, 10.1029/94RG00625.

71. Vicente GA, Davenport JC, and Scofield RA, 2002: The role of orographic and parallax corrections on real time high resolution satellite rainfall rate distribution. Int. J. Remote Sens, 23, 221–230, doi:10.1080/01431160010006935.

72. Adam JC, Clark EA, and Lettenmaier DP, 2006: Correction of Global Precipitation Products for Orographic Effects. J. Clim, 19, 15–38

73. Alvioli M, and Baum RL, 2016, Serial and parallel versions of the Transient Rainfall Infiltration and Grid-Based Regional Slope-Stability Model (TRIGRS): U.S. Geological Survey software release, 10.5066/F73J3B27

74. Zieher T, Rutzinger M, Schneider-Muntau B, Perzl F, Leidinger D, Formayer H, and Geitner C, 2017, Sensitivity analysis and calibration of a dynamic physically based slope stability model: Nat. Hazards Earth Syst. Sci, 17, 971–992, 10.5194/nhess-17-971-2017

75. Beven K. and Freer J: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol, 249, 11–29, 10.1016/S0022-1694(01)00421-8, 2001.

76. Formetta G, Capparelli G, and Versace P: Evaluating performance of simplified physically based models for shallow landslide susceptibility, Hydrol. Earth Syst. Sci, 20, 4585–4603, 10.5194/hess-20-4585-2016, 2016.

77. Mergili M, Fischer J-T, Krenn J, and Pudasaini SP: r.avaflowv1, an advanced open-source computational framework for the propagation and interaction of two-phase mass flows, Geosci. Model Dev, 10, 553–569, 10.5194/gmd-10-553-2017, 2017.

78. Green DM and Swets J,A (1966). Signal detection theory and psychophysics. New York, NY: John Wiley and Sons Inc. ISBN 0–471-32420–5.

79. Botev ZI (2016). "The normal law under linear restrictions: simulation and estimation via minimax tilting". Journal of the Royal Statistical Society, Series B. arXiv:1603.04166. doi:10.1111/rssb.12162.

80. Breslaw JA, 1994 Random sampling from a truncated multivariate normal distribution. Appl. Math. Lett 7(1), pp. 1–6.

81. Damien P, and Walker SG (2001). "Sampling truncated normal, beta, and gamma densities". Journal of Computational and Graphical Statistics. 10 (2): 206–215. doi :10.1198/10618600152627906

82. Anderson MG, 1990 A feasibility study in mathematical modelling of slope hydrology and stability, Report to Geotechnical Control Office Civil Engineering Services Department, Hong Kong.

83. Kirschbaum DB, Adler R, Hong Y, Kumar S, Peters-Lidard C, and Lerner-Lam A. 2012 Advances in landslide nowcasting: evaluation of a global and regional modeling approach, Environmental Earth Sciences, 66, 1683–1696, doi:10.1007/s12665-011-0990-3.

84. Formetta G, Capparelli G, Rigon R, and Versace P. 2014 Physically based landslide susceptibility models with different degree of complexity: calibration and verification, Proceedings of the 7th International Congress on Environmental Modelling and Software, 15–19 June, San Diego, California, USA.

85. Pianosi F, Sarrazin F, Wagener T. (2015), A Matlab toolbox for Global Sensitivity Analysis, Environmental Modelling & Software, 70, 80–85

86. Ling CX, Sheng VS (2011) Class Imbalance Problem In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. Springer, Boston, MA

87. Laloy E, and Vrugt JA (2012), High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing, Water Resources Research, 48, W01526, doi :10.1029/2011WR010608.

88. Vrugt JA, and Laloy E. (2014), Reply to comment by Chu et al. on "High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing", Water Resources Research, 50, 2781–2786, doi :10.1002/2013WR014425.

89. Sadegh M, and Vrugt JA (2014), Approximation Bayesian computation using Markov chain Monte Carlo simulation: DREAM(ABC), Water Resources Research, 50, doi :10.1002/2014WR015386.

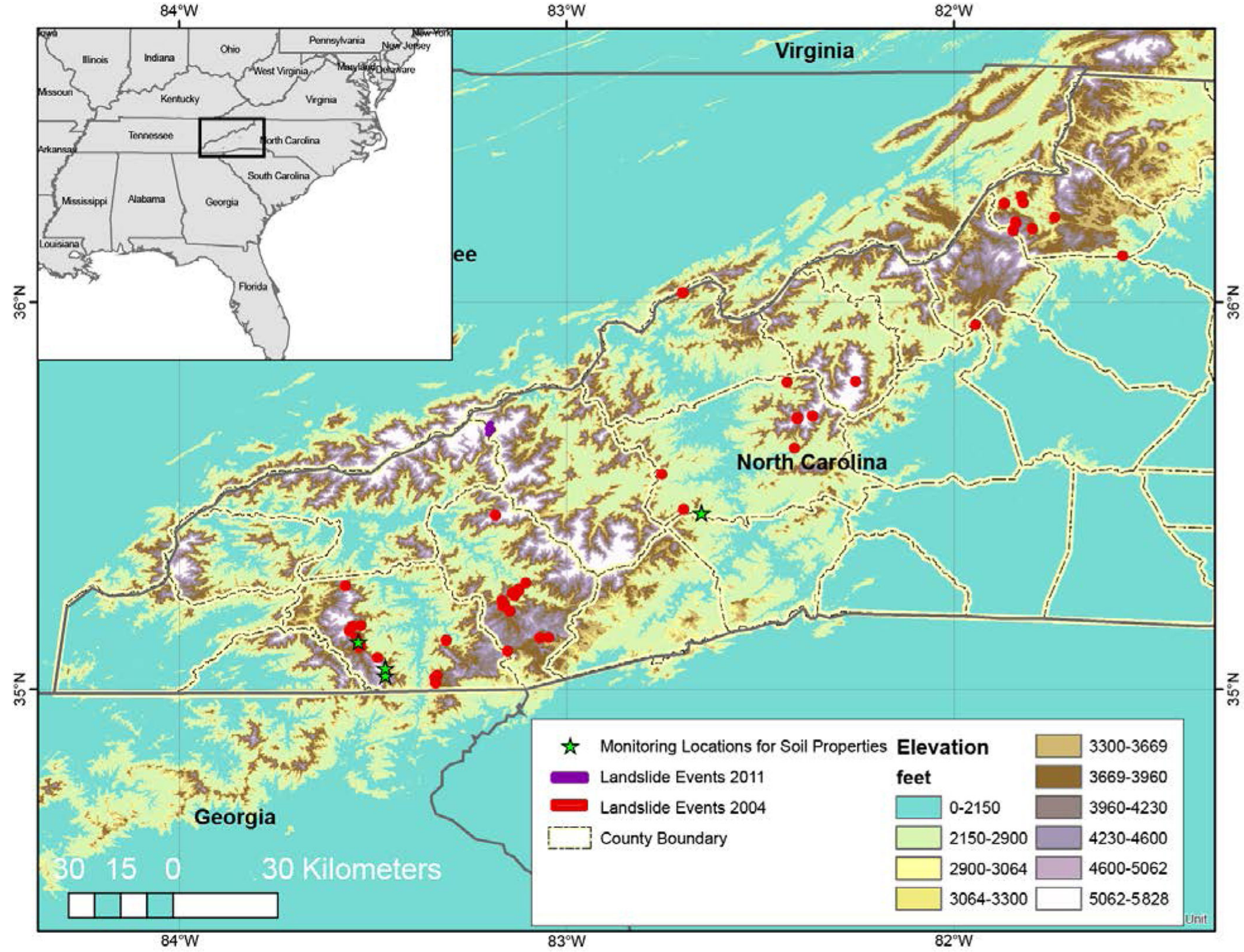

Fig. 1.

Map of North Carolina showing landslide sites considered in this study from the NCGS geodatabase, and the monitoring locations. Elevation data from https://hydrosheds.cr.usgs.gov/dataavail.php.

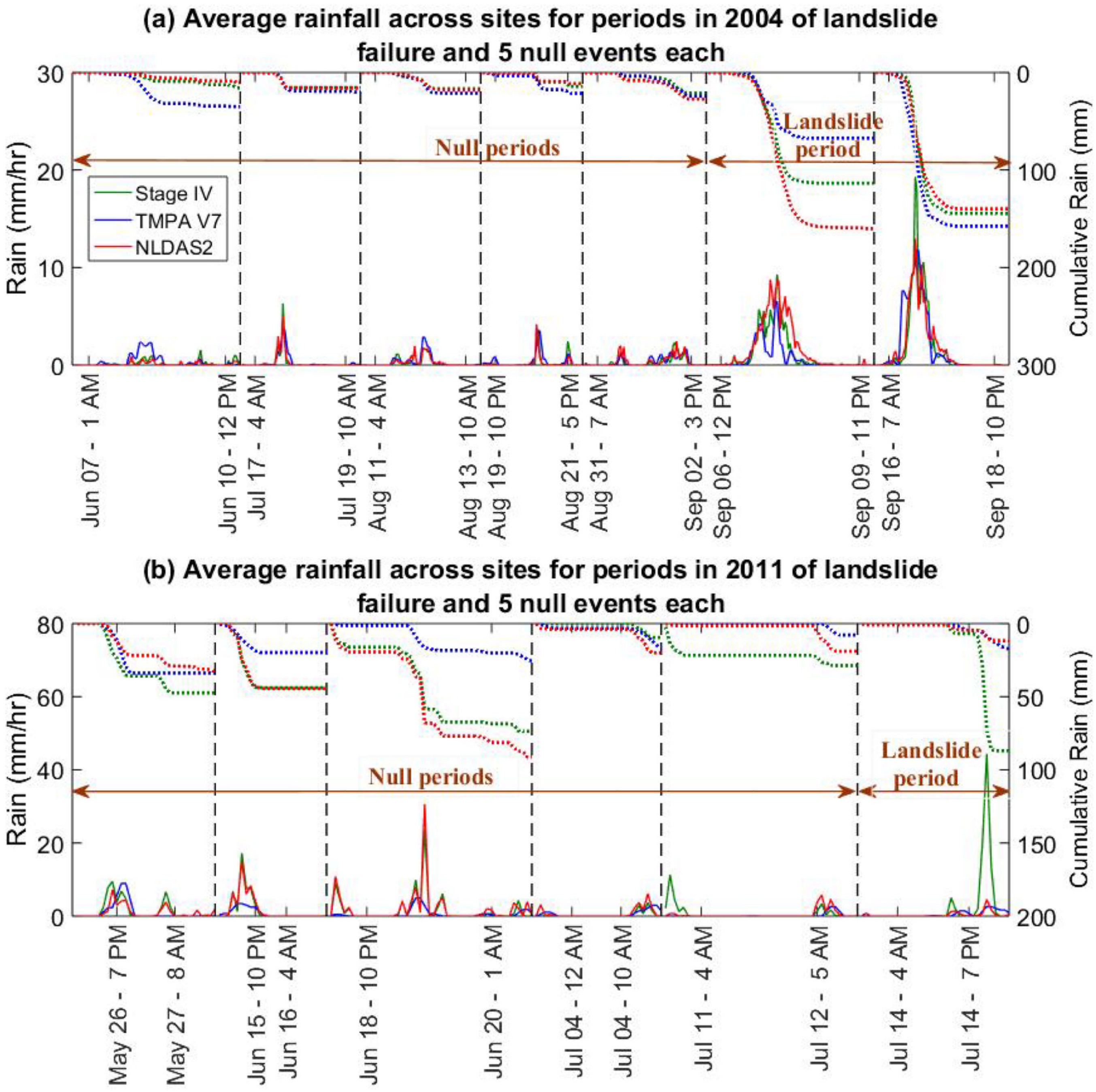

Fig. 2.

Time series of rain (solid lines; left y-axis label) and time-cumulative rain (dotted lines: right inverted y-axis label) averaged across sites during selected null and landslide periods in 2004 (subplot a) & 2011 (subplot b). Subplot b shows possible event-based rainfall errors in the V7 and NLDAS-2 in 2011 reflected by higher rain volumes during their null periods than during the landslide period.

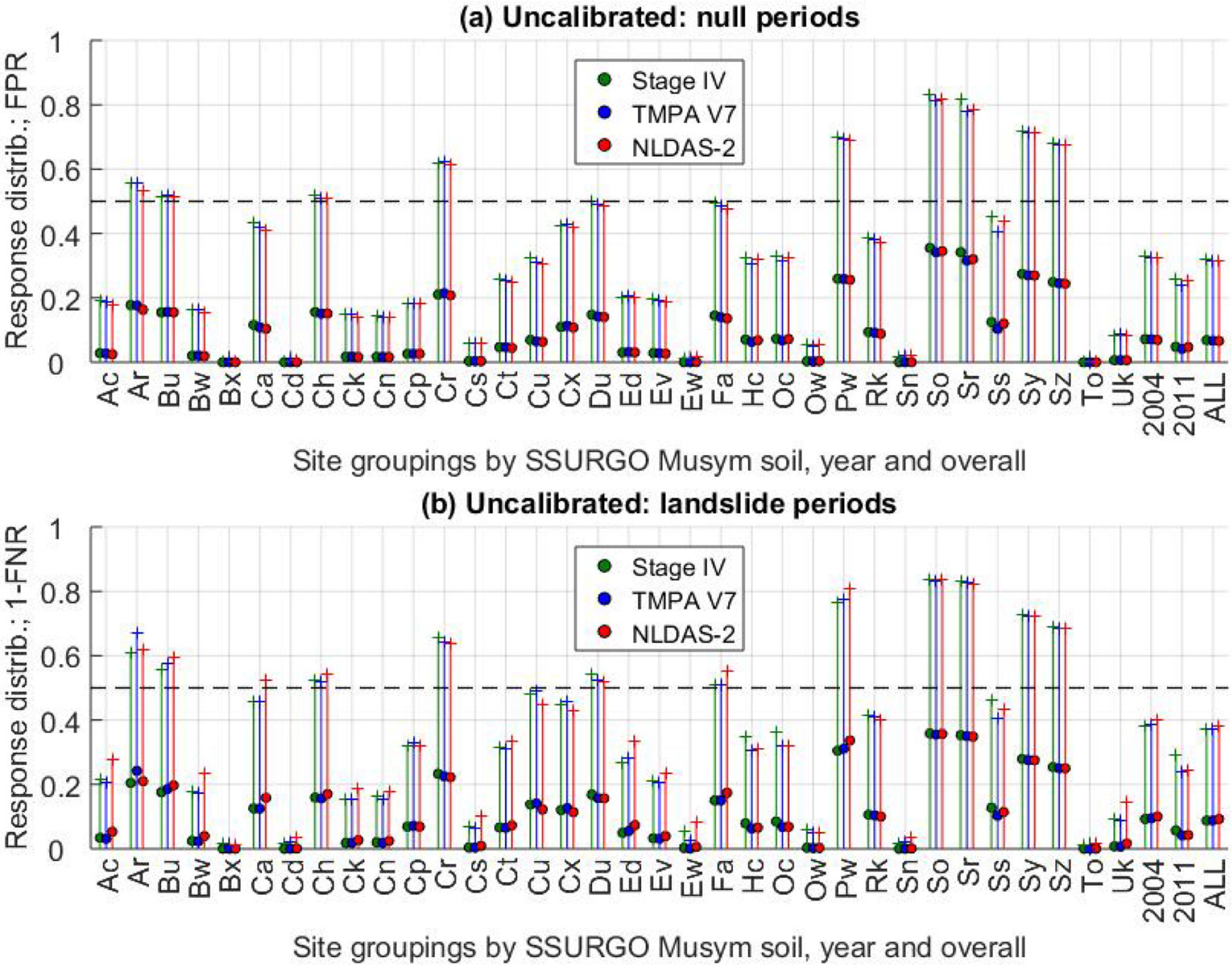

Fig. 3.

Characteristics of mean and standard deviation-defined range for binary response distributions from uncalibrated parameter sets grouped by Musym soils, landslide event year and overall (the latter two form the rightmost three x-axis values). Subplot (a) is for null periods where failure is assumed to not have occurred and so y-axis values close to zero are desirable; subplot (b) is for the recorded slope failure period and so y-axis values close to one are desirable. Circle (see legend) represents the mean for the distribution, vertical line bounded by plus signs represent one standard deviation on either side of mean. Dashed line at the y = 0.5 threshold helps interpret whether any response value (or mean) is closer to 0 or to 1. Per Section 3.5, the means in Subplots (a) are also the False Positive Rates (FPR), and in (b) are also the True Positive Rates (TPR) or the complement of the False Negative Rates (FNR).

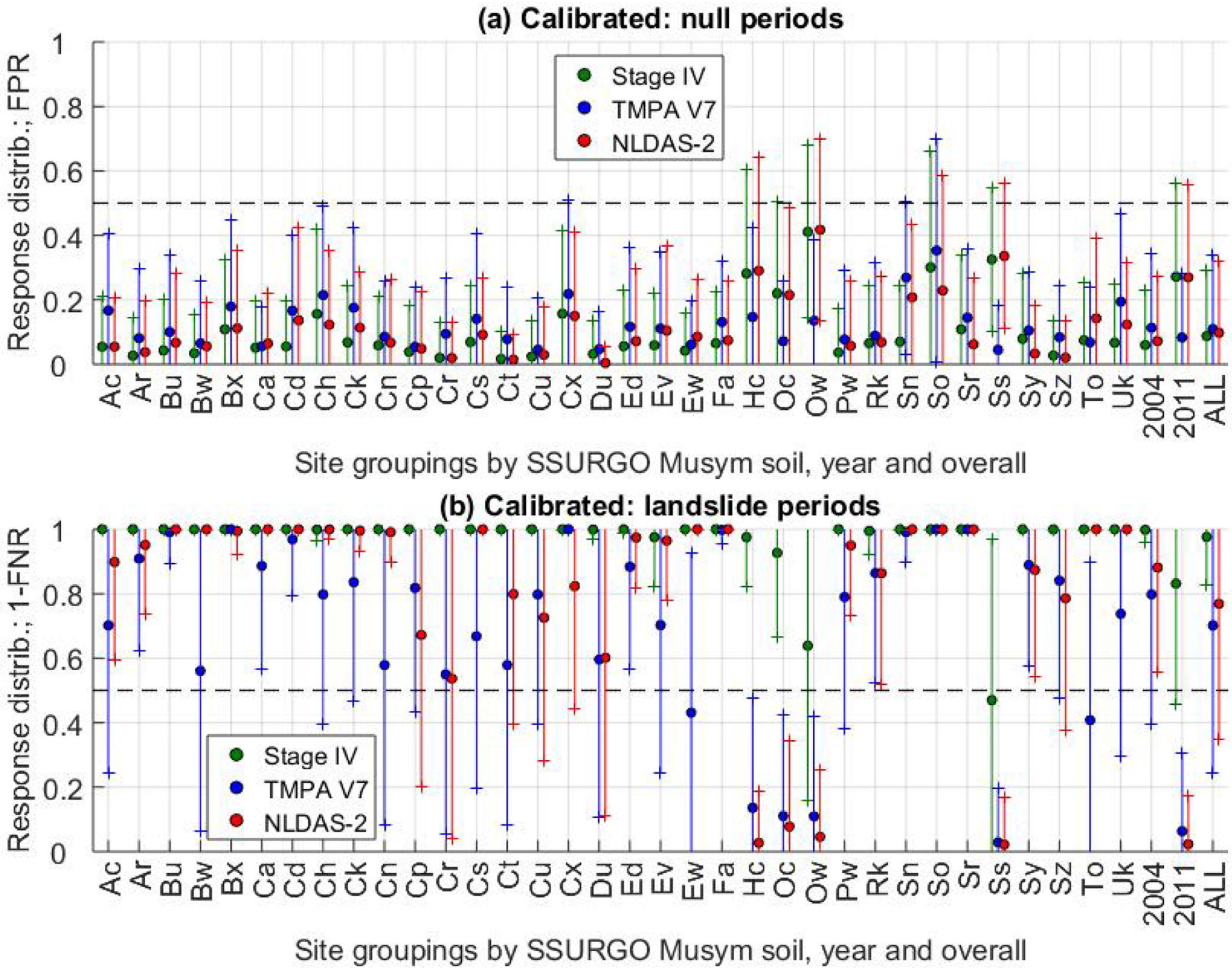

Fig. 4.

As in Fig. 3 but for response distributions from calibrated parameter sets.

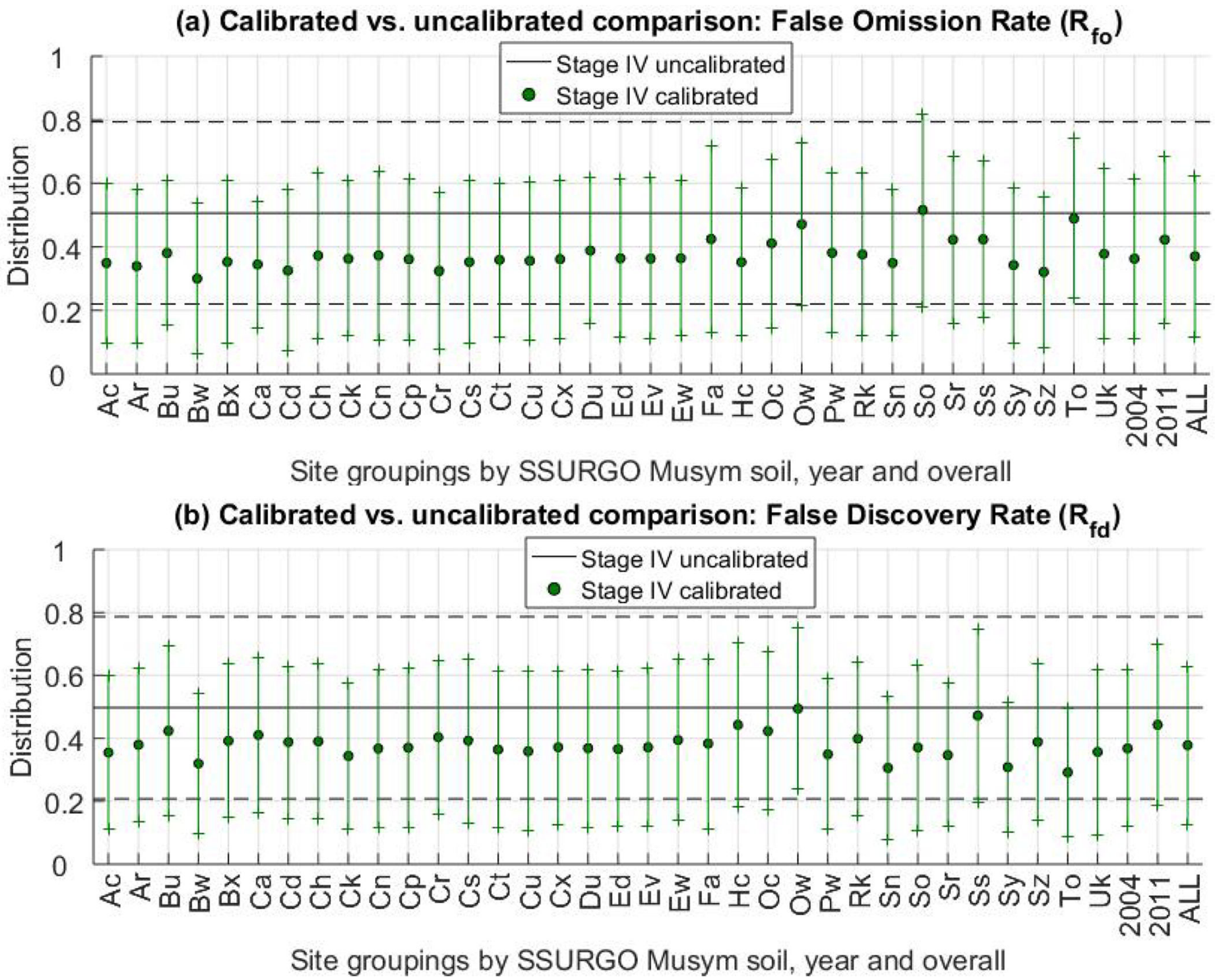

Fig. 5.

Distribution characteristics for hyperparameters R_{fo} (top subplot) and R_{fd} (bottom subplot) grouped by Musym soils, landslide event year and overall (the latter two form the rightmost three x-axis values). Characteristics before calibration are shown by black solid lines for mean and black dashed lines for range bounded by 1σ, while those after calibration are shown by green circles for mean and green plus signs for range bounded by 1σ (σ is the standard deviation).

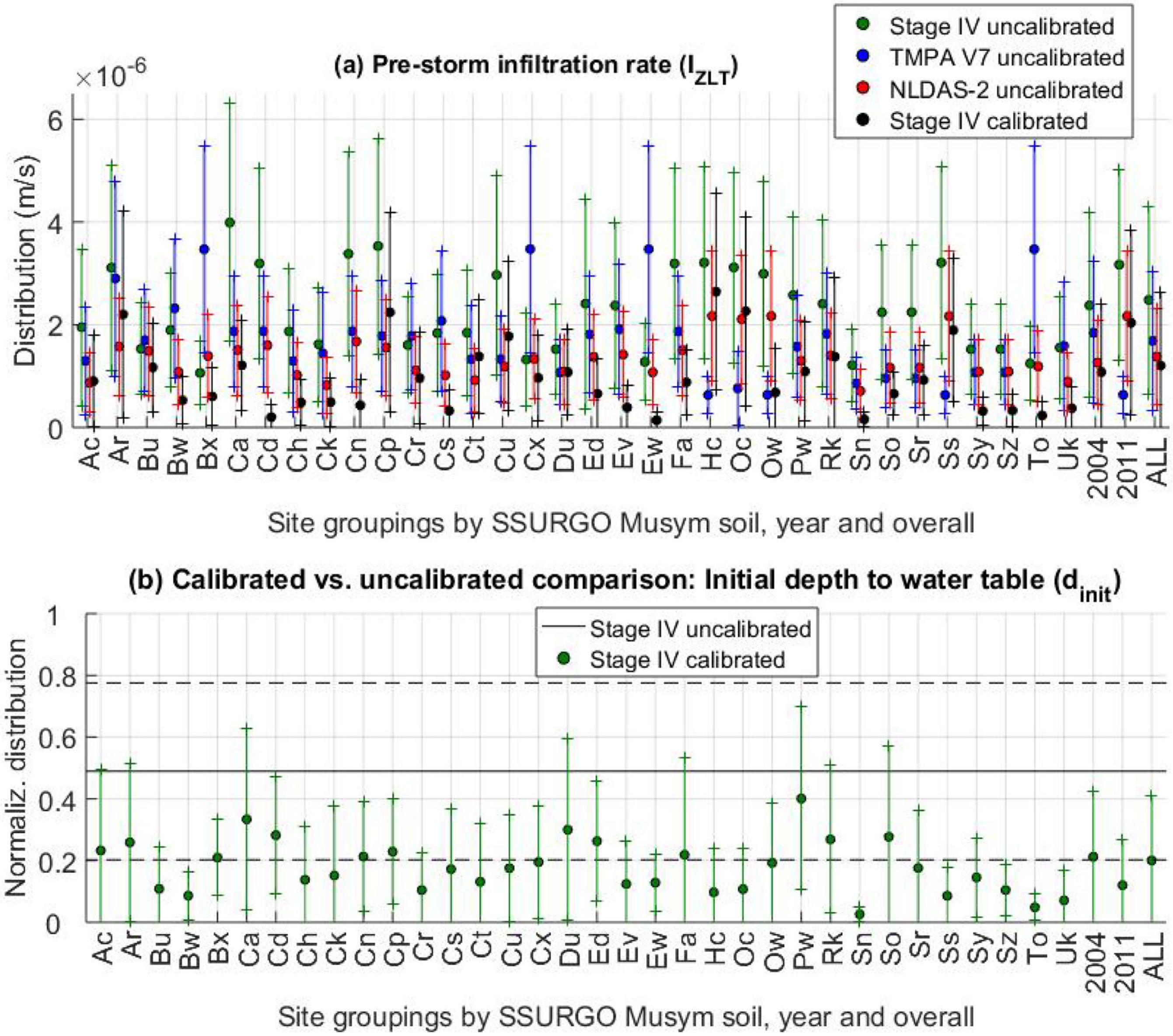

Fig. 6.

Distribution characteristics for initial conditions grouped by Musym soils, landslide event year and overall (the latter two form the rightmost three x-axis values). Subplot (a) shows both the uncalibrated (Stage IV, TMPA V7 and NLDAS-2) and the Stage IV-calibrated I_{ZLT} distribution. Subplot (b) shows the Stage IV calibrated and uncalibrated versions of the normalized d_{init} distribution.

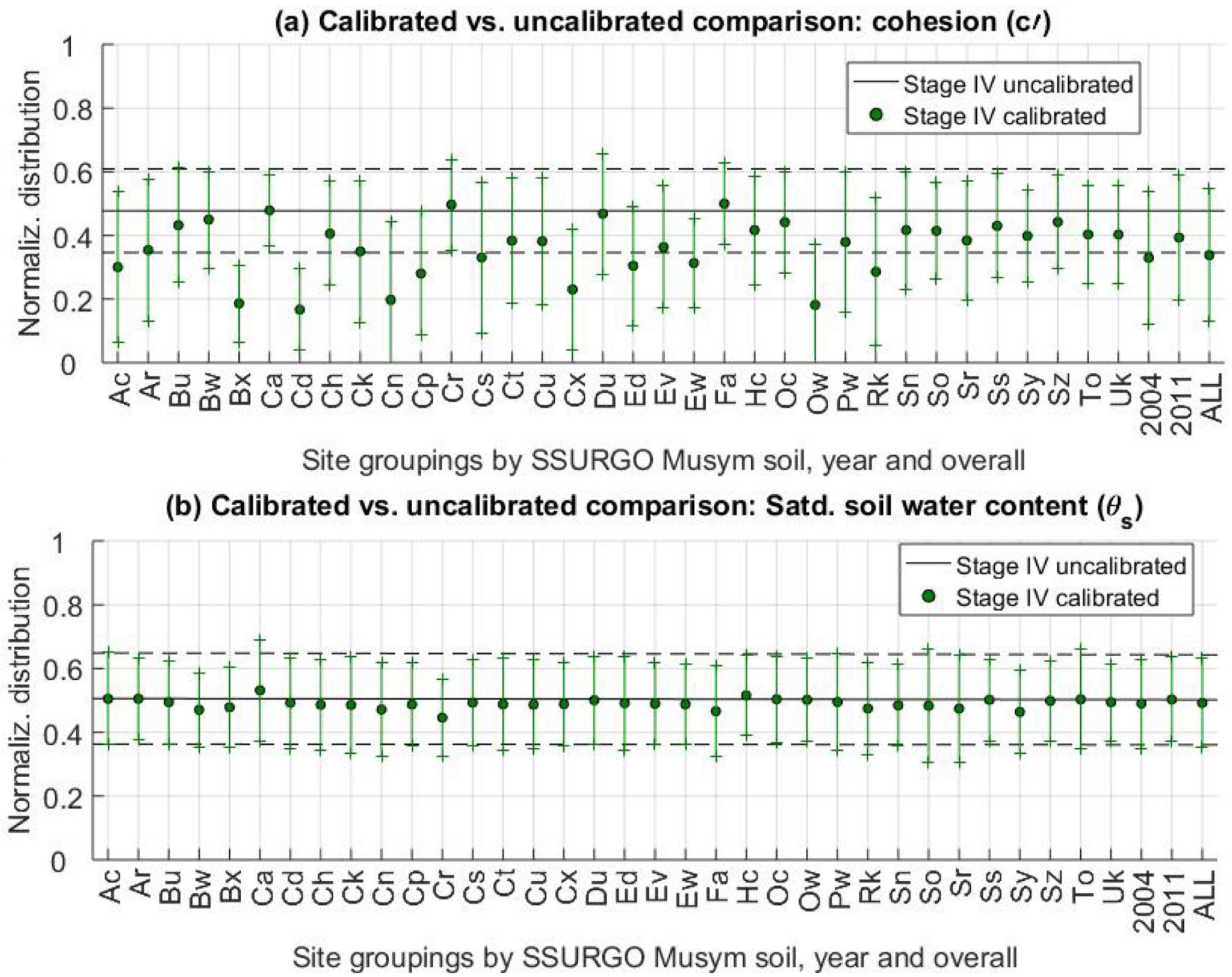

Fig. 7.

Distribution characteristics for example parameters grouped by Musym soils, landslide event year and overall (the latter two form the rightmost three x-axis values). Subplot (a) shows the uncalibrated distribution needing to be shifted toward lower values of effective cohesion (c') for desirable simulations. Subplot (b) shows the distribution almost unchanged from calibrated to uncalibrated distributions for the saturated soil water content (θ_s).

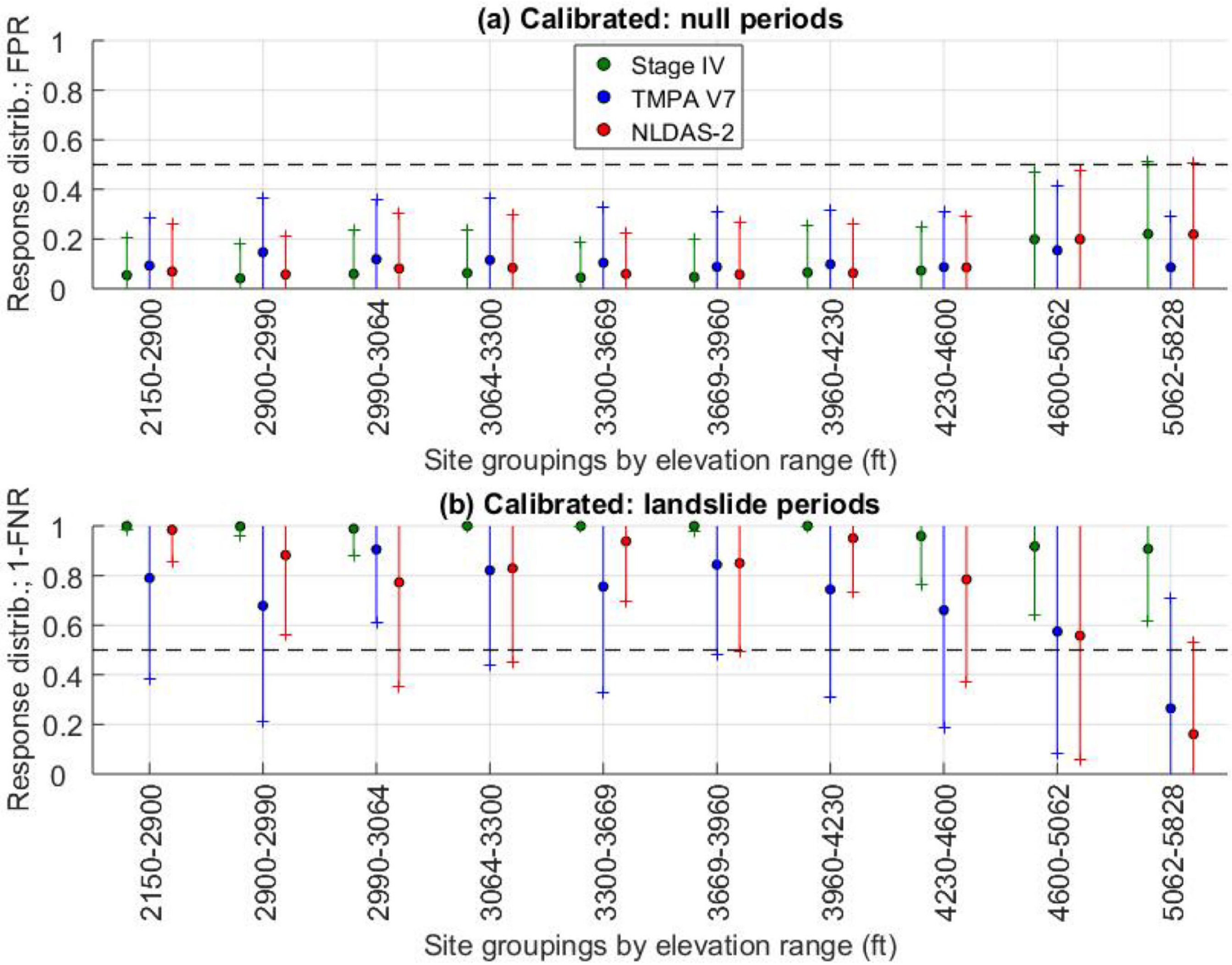

Fig. 8.

As in Fig. 4 but for site groupings by elevation ranges. Subplot (b) shows degradation in modeled response for V7 and NLDAS-2 against Stage IV for the highest elevation ranges.

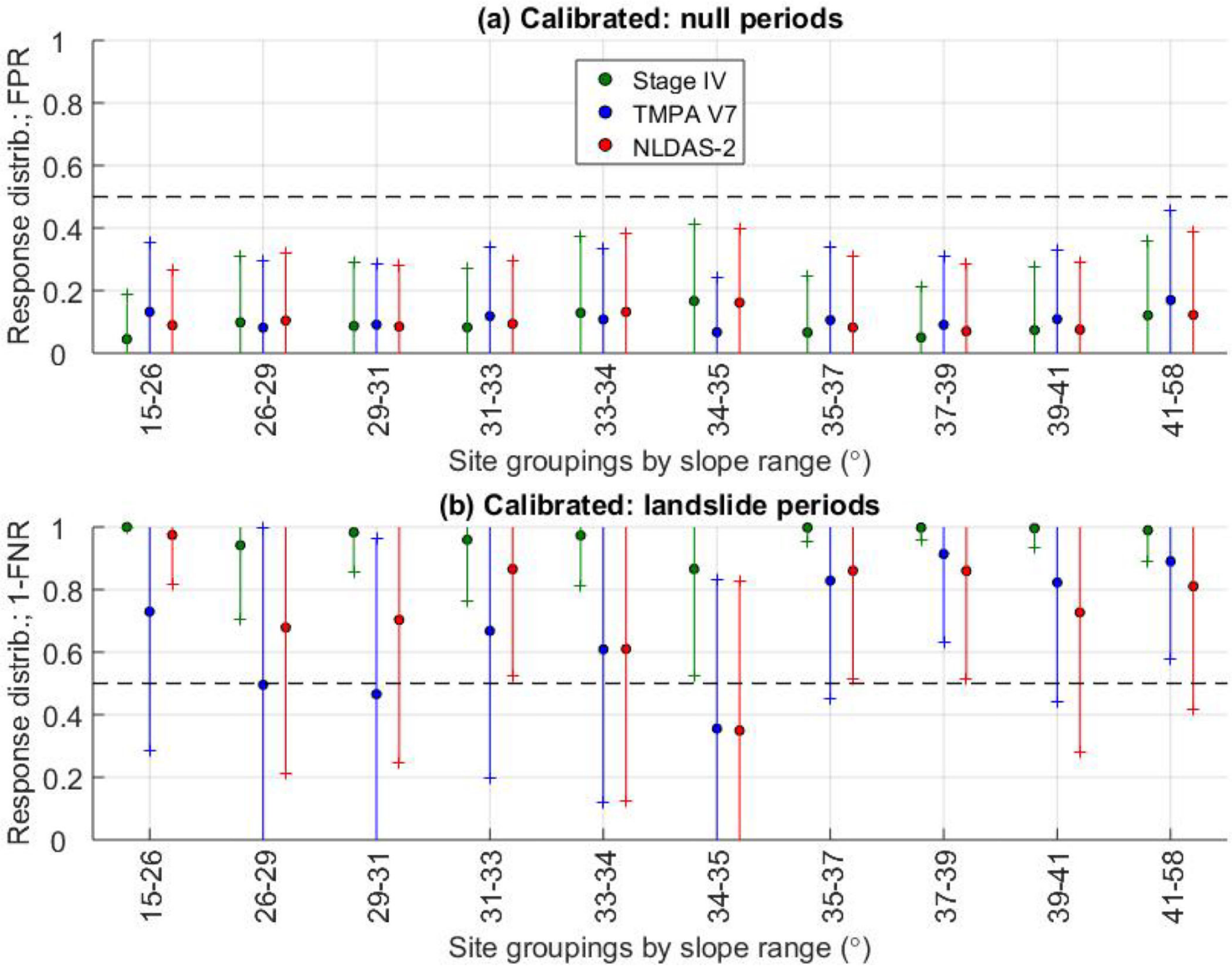

Fig. 9.

As in Fig. 8 but for site groupings by slope ranges.

Table 1

SSURGO Musym soils (ref. Section 2.4) at the NCGS-recorded failure sites used, with the ones also present at the monitoring locations shown in bold.

Musym soil	Descriptive name
Ac	Ashe-Cleveland-Rock outcrop complex, very stony/rocky
Ar	Arkaqua loam, frequently flooded; Ashe-Cleveland-Rock outcrop complex, very bouldery
Bu	Burton-Craggey-Rock outcrop complex, windswept, stony
Bw	Burton-Wayah complex, windswept, stony
Bx	Burton-Craggey-Rock outcrop complex, windswept, very bouldery
Ca	Cashiers gravelly fine sandy loam
Cd	Chandler (gravelly) fine sandy loam; Cataska-Sylco-Rock outcrop complex, very stony
Ch	Cheoah channery loam, stony; Chestnut-Ashe complex, very stony
Ck	Chestoa-Ditney-Rock outcrop complex, very bouldery; Cheoah-Jeffrey complex, very rocky; Chestnut-Edneyville complex, stony
Cn	Chestnut-Edneyville complex, windswept, stony
Cp	**Cleveland-Chestnut-Rock outcrop complex, windswept**
Cr	Craggey-Rock outcrop-Clingman complex, windswept, rubbly
Cs	Cullasaja very cobbly loam, very stony
Ct	Cullasaja very cobbly loam, extremely bouldery; Cullasaja-Tuckasegee complex, very stony;
Cu	**Cullasaja-Tuckasegee complex, stony**
Cx	Craggey-Rock outcrop-Clingman complex, windswept, rubbly
Du	Ditney-Unicoi complex, very rocky
Ed	**Edneyville-Chestnut complex, stony**
Ev	Evard-Cowee complex, stony / moderately eroded
Ew	Evard-Cowee complex, stony
Fa	Fannin fine sandy loam
Hc	Heintooga-Chiltoskie complex, stony
Oc	Oconalufee channery loam
Ow	Oconalufee channery loam, windswept
Pw	**Plott fine sandy loam, stony**
Rk	Rock outcrop-Cleveland complex, windswept
Sn	Saunook loam
So	Soco-Stecoah complex
Sr	Statler loam, rarely flooded; Spivey-Santeetlah complex, stony
Ss	Spivey-Santeetlah-Nowhere complex, very stony; Spivey-Santeetlah complex, very stony
Sy	Sylco-Soco complex, stony
Sz	Sylco-Soco complex, very stony
To	Toecane-Tusquitee complex, boulder
Uk	Unaka-Porters complex, very rocky

Table 2

Uncalibrated parameter distributions. Standard deviation is denoted by σ.

	Parameter	Symbol	Type	Distribution	Source	[a]Bayes Merge?	Bounds
1	saturated soil water content	θ_s	Hydraulic	Normal	ROSETTA	[b]Y	4σ
2	residual soil water content	θ_r		Normal	ROSETTA	[b]Y	4σ
3	vertical satd. hydraulic conductivity	K_s		Lognormal	ROSETTA	[b]Y	4σ
4	specific storage	S_s		Uniform	*Baum et al.* [2010]	[b]N	0.005–0.5 m^{-1}
5	inverse of capillary rise	α		Lognormal	ROSETTA	[b]Y	4σ
6	unit weight of soil	γ_s	Geotechnical	Normal	ROSETTA + NAVFAC DM 7.01	[b]Y	8σ
7	cohesion for effective stress	c'		Normal	ROSETTA + www.geotechdata.info	[b]Y	8σ
8	friction angle	ϕ'		Normal		[b]Y	6σ
9	topographic slope	δ	Geomor phologic	Normal	NCGS, else LiDAR	-	6σ
10	depth to bedrock	Z_{max}		Normal	regression to slope	-	6σ
11	initial depth to water table	d_{init}	Initial condition	Uniform	-	-	$d_{init,F_S=1} - Z_{max}$ or 0−Z_{max} Per eqn [20]
12	pre-storm steady infiltr. Rate	I_{ZLT}		Uniform	rain data	-	0−(0.5 of max. rain {landslide, null periods})
13	false omission rate	R_{fo}	Hyperpa rameter	Uniform	-	-	0–1
14	false discovery rate	R_{fd}		Uniform	-	-	0–1

[a] Bayesian merging with information from monitoring locations

[b] Y/N: Yes/No

Table 3:

Ranges used in this study for geotechnical parameters over soil texture classes at the landslide sites. Missing property range information for texture types in the NAVFAC DM 7.01 or www.geotechdata.info sources have italics font and are filled in or generated here as union of available property ranges for neighboring soil texture classes in the soil texture triangle. Bottom row range gives the assumed equivalent in terms of the standard deviation for the normal distribution.

Soil texture class ▼	**γ_s range (lb/ft^3)**	**c' range (Pa)**	**ϕ' range (°)**
Loam	*81 – 147*	10,000 – 20,000	28 – 32
Loamy sand	*84 – 148*	10,000 – 20,000	31 – 34
Sandy loam	*84 – 148*	*10,000 – 20,000*	*25 – 34*
Normal distribution range ►	6σ	6σ	4σ

Note: Ranges in italics are generated as union of ranges of neighboring soil texture types in the soil textural triangle. E.g., the ϕ′ value range of 25–34° for sandy loam covers the ϕ′ value ranges for loamy sand (31–34°), sandy clay loam (31–34°), loam (28–32°) and silty loam (25–32°).