JGR Earth Surface

RESEARCH ARTICLE
10.1029/2018JF004961

This article is a companion to Barnhart et al. (2020b, 2020c), https://doi.org/10.1029/2018JF004963 and https://doi.org/10.1029/2019JF005287.

Key Points:
- Multiple landscape evolution models are developed for application in a well-constrained site undergoing locally rapid (4 m/ka) erosion
- Global sensitivity analysis reveals that uncertainty in initial and boundary conditions has low impact on model output metrics
- Global sensitivity analysis finds that channel erosion parameters dominate and that select metrics are sensitive to hillslope diffusivity

Supporting Information:
- Supporting Information S1

Correspondence to:
K. R. Barnhart,
katherine.barnhart@colorado.edu

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

Received 26 NOV 2018
Accepted 11 FEB 2020
Accepted article online 18 FEB 2020

Inverting Topography for Landscape Evolution Model Process Representation: 1. Conceptualization and Sensitivity Analysis

Katherine R. Barnhart[1,2], Gregory E. Tucker[1,2], Sandra G. Doty[3], Charles M. Shobe[1,2,4], Rachel C. Glade[1,5,6], Matthew W. Rossi[7], and Mary C. Hill[8]

[1]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA, [2]Department of Geological Sciences, University of Colorado Boulder, Boulder, CO, USA, [3]Denver, CO, USA, [4]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany, [5]Institute for Arctic and Alpine Research, University of Colorado Boulder, Boulder, CO, USA, [6]Earth and Environmental Sciences Division, Los Alamos National Lab, Los Alamos, NM, USA, [7]Earth Lab, University of Colorado Boulder, Boulder, CO, USA, [8]Department of Geology, University of Kansas, Lawrence, KS, USA

Abstract Despite considerable community effort, there is no general set of equations to model long-term landscape evolution. In order to determine a suitable set of landscape evolution process laws for a site where postglacial erosion has incised valleys up to 50 m deep, we generate a set of alternative models and perform a multimodel analysis. The most basic model we consider includes stream power channel incision, uniform lithology, hillslope transport by linear diffusion, and surface-water discharge proportional to drainage area. We systematically add one, two, or three elements of complexity to this model from one of four categories: hillslope processes, channel processes, surface hydrology, and representation of geologic materials. We apply methods of formal model analysis to the 37 alternative models. The global Method of Morris sensitivity analysis method is used to identify model input parameters that most and least strongly influence model outputs. Only a few parameters are identified as important, and this finding is consistent across two alternative model outputs: one based on a collection of topographic metrics and one that uses an objective function based on a topographic difference. Parameters that control channel erosion are consistently important, while hillslope diffusivity is important for only select model outputs. Uncertainty in initial and boundary conditions is associated with low sensitivity. Sensitivity analysis provides insight to model dynamics and is a critical step in using model analysis for mechanistic hypothesis testing in landscape evolution theory.

1. Introduction

A fundamental question in geomorphology is the extent to which we can infer process from form. Topographic form clearly contains information—a dune field, for example, appears fundamentally different from a deglaciated alpine valley. At a finer scale, however, the extent to which topography or other observables at Earth's surface can be used to infer the functional structure of the governing equations that describe the evolution of the land surface is an open question. Further, the answer is likely to vary depending on environmental conditions, constraint on initial conditions, the role of autogenic processes, and the timescale of interest. The defining, testing, and improving of these governing equations, and numerical models built to solve them, are core goals of quantitative geomorphology. They serve to formalize the current state of knowledge, are used to interrogate the dynamical properties of Earth's surface, and allow us to make predictions with quantified uncertainty.

Here we confront the challenge of needing predictions for a system in which neither the physical processes nor the parameter values for a sufficient model are known. We do this at a site where hazardous waste storage necessitates predictions of potential long-term future erosion in support of management decisions. To make successful predictions of future erosion, it is essential to first evaluate the success of landscape evolution models in predicting modern topography given plausible model parameters and known past environmental conditions. The watershed we assess in this study (the "West Valley Site") is located in western New York State, USA, in the Cattaraugus Creek basin, which drains to Lake Erie (Figure 1). At the site, a

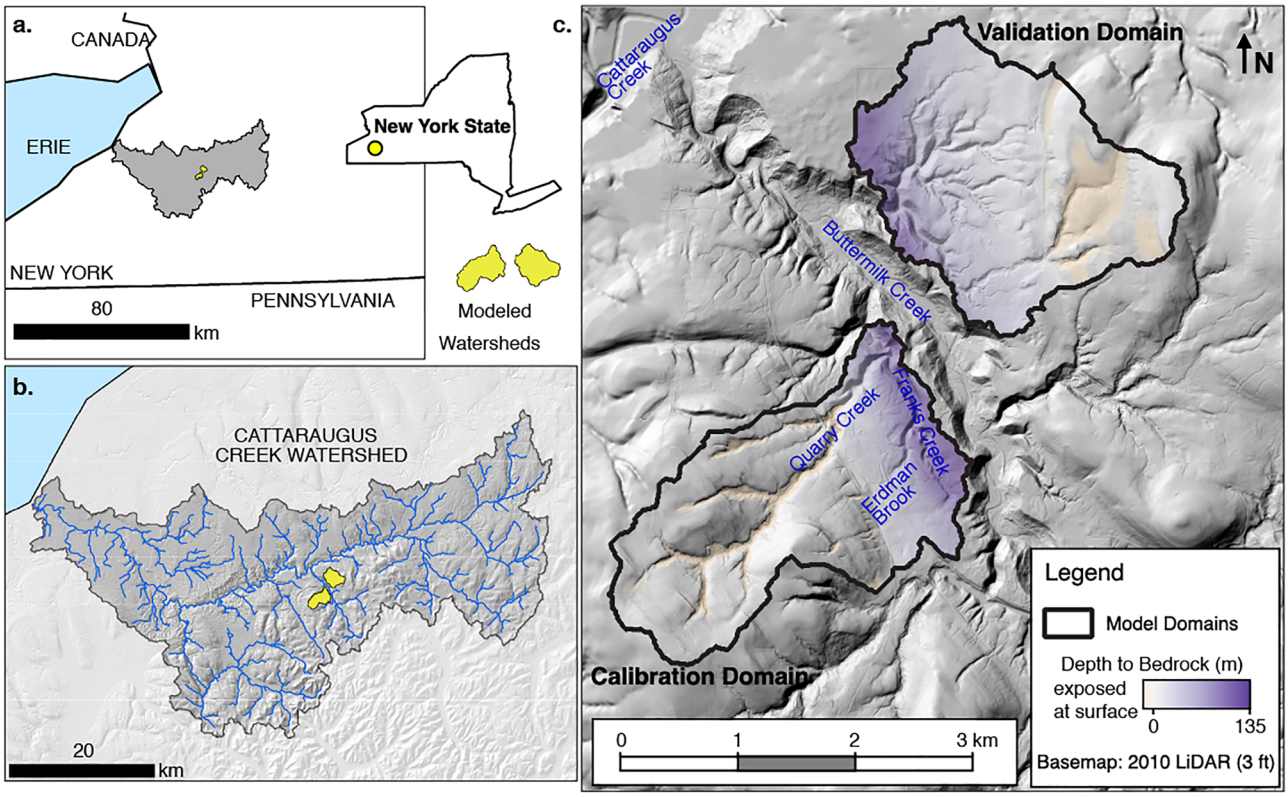

Figure 1. Location of calibration and validation domains in New York State (a) and the Cattaragus Creek Watershed (b). (c) The modern site topography and depth to bedrock within each of the two domains (see supporting information Text S3.2 for source of depth to bedrock). Names of hydrographic features are shown in blue text.

stream network has incised up to 50 m in the last ~13 ka into a glacial till plateau (Wilson & Young, 2018). The larger drainage basin containing the study watershed has a well-constrained postglacial surface, and the watershed outlet downcutting history has been estimated from dating of terraces and paleosurfaces. The site is thus a natural experiment that is well-suited to the calibration and testing of multiple alternative models (Tucker, 2009). To this end, we develop and apply 37 alternative models of landscape evolution in a consistent framework of sensitivity analysis, calibration, and validation to identify which models are able to capture the evolution of the site from glacial retreat to the present. The alternative models represent a hierarchical design in which a simple model systematically gains elements of complexity. Comparing these models thus serves as a formalized hypothesis test of landscape evolution theory.

While the problem of multimodel analysis has been considered in other fields of environmental science (e.g., Clark et al., 2008; Foglia et al., 2013), in this work we present the most extensive use of multimodel analysis and model calibration in geomorphology to date. We present both a case study based at the West Valley site, and a workflow for applying these techniques elsewhere. We outline a philosophy of how these techniques can be used in geomorphology (section 2), place our work in the context of prior applications of model analysis methods (section 2.4), and discuss lessons about application of these methods in the future. In sections 4 and 7, we focus on defining a model set and the insights gained from a comprehensive sensitivity analysis. In a companion paper (Barnhart et al., 2020b), we present a calibration and validation of the set of alternative landscape evolution models. The appropriate ranges of parameter values for common geomorphic transport laws, which are used in both sensitivity analysis and calibration, are discussed a second companion paper (Barnhart et al., 2020c) and in the supporting information.

2. Model Analysis Framework

We use model analysis to quantitatively assess landscape evolution models. Before describing our particular application, we summarize model analysis methods and their limitations, describe the philosophy behind our approach, place the work in the context of similar applications, and outline our strategy.

2.1. Philosophy

We begin by defining what we mean by a *model*. A model, **F**, quantitatively links conceptual and theoretical information about a study system with defined input parameters, **m**, and model outputs, **o**, that represent *simulated equivalents* of observations, **d**:

$$\mathbf{F}(\mathbf{m}) = \mathbf{o} \simeq \mathbf{d}\,. \tag{1}$$

By this definition, a model may be a set of equations with a closed-form solution, or in the case of models presented here, equations that require inverse methods in which many forward models are evaluated in order to estimate **m** such that difference between model and data is minimized. Inputs are all specified before a model run, and may either be continuous (such as an erosion rate coefficient) or categorical (such as the selection of one among several potential initial topographic configurations). The $\mathbf{o} \simeq \mathbf{d}$ in equation (1) acknowledges that the relationship between observations and simulated equivalents may not be exact.

In this work, we consider categorical variations in a model's governing equations as the construction of a different, *alternative* model. A set of governing equations represents a hypothesis for the dynamics of a model's state variables. Thus a set of alternative models, each with different governing equations, represents a set of alternative hypotheses for a system. We consider one model to be distinct from another if it uses a different mathematical form to describe the relationship between an independent and a dependent variable. By this definition, for example, the Manning and Chezy relationships for flow velocity and water depth in open-channel flow would be considered distinct models, because they apply different exponents to the relationship between velocity and depth (2/3 and 1/2, respectively). One implication of our definition of model is that a single computer program can potentially support the use of more than one model. For example, we would consider changing the sediment transport formula from Einstein to Wilcock-Crowe within the CAESAR-Lisflood numerical model—as is done in Skinner et al. (2018)—as representing the construction of two alternative models (Einstein, 1950; Wilcock & Crowe, 2003). We also consider two models to be distinct if they vary in the number of degrees of freedom or parameters. An implication is that two models can be distinct even if one represents a superset of the other; for example, we would consider $y = mx$ and $y = mx + b$ to be different models, even though the first is a subset of the second, because they differ in the number of parameters required. This definition permits us to construct hierarchical model sets, in which fixing a parameter value results in recovering a simpler model. These are useful for formally testing whether adding an additional degree of freedom to a model (that is, increasing its complexity) improves simulation performance. We note that there is no consensus on the definition of a *distinct model*; others have, for example, treated changes in governing equations as categorical changes to model inputs (e.g., Pianosi et al., 2016; Skinner et al., 2018). We revisit this issue at the end of section 2.3.

State variables or other quantities simulated by a model are compared with observational data and are referred to as *simulated equivalents*. Input parameters may be specified using measured values, or estimated based on calibration. As with many applications in geomorphology, our primary observation and the main state variable evolved by our models is topographic elevation. However, depending on the application, other derived metrics or targeted geographic locations may be more important to evaluating model success than direct comparison of elevations between a model and reality, a topic we address in sections 2.3 and 6.

The methods and statistical tools of model analysis are used to assess model dynamics, simulate system performance, and quantify the relationship between uncertain inputs and uncertain predictions. Common methods in model analysis include sensitivity analysis, calibration, and validation. They are described briefly here and more extensively in the subsequent section. Sensitivity analysis addresses how parameter values (**m**) influence model outputs (**o**), and it is often used to identify the most important input parameters in a particular application. Calibration identifies the values of **m** that minimize the misfit between simulated and observed equivalents (e.g., with a sum of squared residuals: $||\mathbf{F}(\mathbf{m}) - \mathbf{d}||^2$). This results in an expected value and uncertainty estimate for each parameter associated with the "best" version of a model. Validation applies a calibrated model to a different data set, and assesses the model's ability to reproduce it.

In many geomorphic systems it is not clear a priori which model is both necessary and sufficient to simulate a system. Suppose for example that one starts with a relatively simple model that, given initial guesses as to its governing parameters, fails to explain the observations of interest. Improving this simple model may only require adjusting one or more parameter values, or it may require more fundamental changes in

model structure. Structural changes can include evolving additional (or different) state variables, using different governing equations, or modifying a model in some other categorical way. In the context of landscape evolution modeling, which commonly couples together models for surface hydrology, hillslope sediment transport, and erosion by surface water, a common structural change is the use of a different *process law*. An example of a choice among process laws is the use a linear or nonlinear diffusion rule for the calculation of hillslope sediment flux (e.g., Roe et al., 2008). A landscape evolution model can be made more complex by using process laws that require more calibrated parameters, or by adding additional process laws, such as a weathering rule to calculate the thickness of a dynamic soil layer.

In order to determine which model choices improve model performance, we use *multimodel analysis* (Burnham & Anderson, 2003) in which alternative **F**s are considered. Under this method, we apply sensitivity analysis, calibration, and validation on multiple landscape evolution models in order to compare then within a consistent framework. By comparing alternative models calibrated using the same method and against the same data, we are able to ask "which model performs best?" and to quantify model structure uncertainty (Burnham & Anderson, 2003; Foglia et al., 2013; Poeter & Hill, 2007). This approach is not fully objective, as the choice of models and the choice of simulated equivalents are themselves *choices* (see section 2.3 for further discussion of limitations). However, the key advantage of the multimodel approach is that it transparently sets the model space and makes explicit the basis for model performance and comparison.

2.2. Model Analysis Methods

The core model analysis methods applied here and in Barnhart et al., (2020b) are sensitivity analysis, calibration, and multimodel analysis. Sensitivity analysis typically serves one or more of the following three purposes (Jansen, 1999; Pianosi et al., 2016; Saltelli et al., 2008):

1. *ranking* input parameters in their relative influence on variability of output values,
2. *screening* input parameters that have low influence on the variability of output values,
3. *mapping* regions of parameter space that produce outlier output values.

There are two main classes of sensitivity analysis: *local* approaches, which focus on the relationship between inputs and outputs at one point in parameter space, and *global* approaches, which consider the entire parameter space (Pianosi et al., 2016; Saltelli et al., 2008). Local methods are computationally frugal and require definition of a central point. Global methods require many more model evaluations, and the definition of valid parameter space. If the values provided for either the central point or the parameter ranges are poorly constrained, this influences the results. Sensitivity analysis is commonly used to understand model dynamics, identify and screen out unimportant parameters in advance of calibration, and relate input and output uncertainties (Jansen, 1999; Pianosi et al., 2016; Saltelli et al., 2008). As described in section 7, in our application we focus on understanding dynamics and screening. See Saltelli et al. (2008) or Pianosi et al. (2016) for more extensive reviews of sensitivity analysis.

Calibration (also called inversion or parameter estimation) seeks to identify the parameter set that minimizes the misfit between model and observations (e.g., Hill & Tiedeman, 2007; Tarantola, 1987; Tarantola & Valette, 1982). Given a defined *objective function* or formal statement of model performance (often the sum of squared weighted residuals between observations and simulated equivalents), calibration seeks to find the *global minimum* in a parameter hyperspace with a number of dimensions equal to the number of estimated parameters. There are two major classes of calibration algorithms. *Gradient-based methods* use the gradient and or higher order information about the objective function surface to search for a minimum. In contrast, *global methods* take a sampling approach to identify a global minimum, and may include Bayesian calibration or evolutionary algorithms (see Adams et al., 2017a, for additional methods).

Finally, multimodel analysis and model selection compares alternative calibrated models and ranks their relative performance (Burnham & Anderson, 2003; Poeter & Hill, 2007). These methods can be used to draw an inference about which alternative model provides a sufficient description of the system in question. They can also be used to assess model structure uncertainty. If we consider alternative governing equations (and thus models) as hypotheses for the minimally sufficient representation of a system, multimodel analysis permits formal hypothesis testing. The comparison of models is typically done using performance statistics that incorporate the objective function of a calibrated model and a penalty based on the number of model parameters (e.g., Akaike's information criterion; Akaike, 1974).

Given this background in model analysis methods, a reasonable basis emerges for considering model governing equations as different models (as is done in this work) or as categorical parameter values (as is done in Pianosi et al., 2016 and Skinner et al., 2018). If the goal of the work is to understand model sensitivity only, then considering changes in governing equations as categorical parameter changes provides insight. Under this case, a practitioner must address the potential issue of different governing equations resulting in variable numbers of continuous parameters. If instead the analysis aims to use calibrated versions of alternative models for hypothesis testing or multimodel uncertainty analysis, then it is appropriate to consider each set of governing equations as a *distinct model*.

2.3. Limitations and Challenges

Individual methods of model analysis (such as the Method of Morris described in more detail in section 7, or the Gauss-Newton gradient-based calibration used in Barnhart et al., 2020b) each have limitations associated with their algorithms. For example, the results of many sensitivity analysis methods may be sensitive themselves to the extent of defined parameter space, and gradient-based calibration methods are not guaranteed to find a global minimum but may instead get trapped in a local minima. Enumerating these method-specific limitations is beyond the scope of this contribution. However, there are a number of general limitations to the application of formal model analysis to landscape evolution modeling that merit discussion.

First, we restate the primary benefit of formal model analysis: input parameters and their ranges, definition of simulated equivalents, and alternative model structures are all formally defined. A model is identified as "good" on the basis of its objective function, rather than on the basis of expert judgment.

Challenges in applying model analysis to landscape evolution and Earth surface dynamics models typically fall into the following three categories:

1. When do calibration and observation time periods coincide?
2. Are model physics and calibration time periods compatible?
3. What state variables (or statistics thereof) have simulated equivalents?

Landscape evolution and Earth surface dynamics models are typically designed for the simulation of Earth's surface on timescales that range from years to geologic epochs (10^6+ years). However, the modern observational record is much shorter. The minimal overlap between simulation and observation timescales often manifests in a model evaluation that compares end-of-simulation with modern topography (e.g., Hancock et al., 2002, 2010). This contrasts with many model analysis studies in surface and groundwater hydrology, which evaluate model performance based on matching observed and modeled time series of well head or stream discharge. Assessing simulations at time points in the geologic past requires that simulated equivalents exist in paleoclimate or sedimentary records.

A second issue related to time is the timescale assumed in deriving a model's governing equations. Many landscape evolution models assume an effective discharge (see, e.g., review by Tucker & Hancock, 2010). These models may not be appropriate to apply on a storm event timescale as they do not represent the physics operating on those timescales. Ultimately whether a model is appropriate to apply to a particular case is an empirical question of model validation, and the applicability is likely to depend on what model outputs are needed. For example, simulation of a single storm event with a model that assumes a steady discharge in its derivation may successfully simulate total eroded volume but fail at simulating spatially distributed patterns of erosion and deposition.

Finally, there is no clear consensus on the most appropriate simulated equivalents with which to evaluate model performance (Hancock & Willgoose, 2001; Hancock et al., 2010, 2011; Howard & Tierney, 2012; Ibbitt et al., 1999; Perera & Willgoose, 1998; Skinner et al., 2018). For example, landscape evolution models typically evolve topography as a state variable. Comparison between modern high resolution topography (e.g., Passalacqua et al., 2015) and end-of-simulation topography seems an attractive and promising option. Landscape models are known to demonstrate sensitivity to initial topography when the initial surface has low relief (Hancock, 2006; Hancock et al., 2016; Ijjász-Vásquez et al., 1992; Kwang & Parker, 2019; Willgoose et al., 2003). The initial condition for a geomorphic model, and in particular the topography at a given time in the past, is usually unknown. Special cases in which the geomorphic setting provides clues that allow reconstruction of paleotopography are therefore especially useful as natural experiments (Tucker, 2009). The degree of model sensitivity to initial topography, and thus the applicability of a one-to-one mapping between

an observed and simulated elevation field, should in principle depend both on the magnitude of uncertainty in paleotopographic reconstruction, and on the potential for drainage pattern rearrangement.

Some of the terrain metrics that have been used to evaluate landscape model performance are functions rather than scalars. Examples include the area-slope relationship, the width function, the hypsometric curve, and the cumulative area distribution (Hancock et al., 2002, 2010). This presents a challenge as many standard model analysis methods are designed for scalar outputs. Reformulating these analyses into scalars (e.g., by the nonparametric Kolmogorov-Smirnov test statistic comparing simulated and observed distributions) may prove useful. Further, as we report in a companion paper (Barnhart et al., 2020b), once a set of simulated equivalents based on scalar terrain statistics has been identified, defining the weights used to combine them into an objective function is a nontrivial task (see Hill & Tiedeman, 2007, Guideline 6 for discussion of weights).

Construction of an objective function need not be limited to a single state variable. Furbish (2003) explored the additional discriminatory power of a coupled model of topography and soil depth. Tests of the SIBERIA landscape evolution model have incorporated erosion pin and ^{137}Cs observations (Hancock et al., 2011, 2018). This approach, however, requires models that represent additional state variables, and observational data on those variables.

The advent of high resolution topography data provides the process geomorphology community with new data sets that are spatially rich but temporally sparse (often limited to just a single snapshot in time). While the development of robust methods for deriving process information from high-resolution topography is still an ongoing research area (e.g., how to best combine topography data with other observation types, how to statistically reduce topographic data), this data source has an additional benefit: we can *see* what the most important state variable should look like. The availability of high-spatial resolution images of a key state variable is an advantage that other environmental modeling disciplines, such as groundwater hydrology, lack. The accessibility of topographic data is one reason that many studies have combined quantitative outputs with qualitative comparisons between model output to assess model performance (e.g., Coulthard & Skinner, 2016; Hancock & Coulthard, 2012; Hancock et al., 2010, 2015).

Taken in whole, the limitations associated with long time scales, poor temporal resolution, and general lack of data other than topography are not fatal, but are symptoms of a nascent research area. Through iterative use of model analysis, identification of persistently good objective functions (e.g., Krause et al., 2005) is likely.

2.4. Model Analysis in Landscape Evolution Processes

Most applications of sensitivity analysis in landscape evolution process modeling focus on ranking and screening, typically with the aim of understanding model dynamics. As described in Skinner et al. (2018), exploratory sensitivity analysis tends to be more common than formal sensitivity analysis in hydrology and geomorphology. Where formal sensitivity analysis has been applied, the Method of Morris (MoM, used here and discussed further in section 7; Morris, 1991) is commonly used.

Two applications of formal sensitivity analysis in landscape evolution modeling include applications of MoM to the CAESAR model (Coulthard et al., 1998, 2002) in the context of the Tagliamento River in Italy on a subdecadal timescale (Ziliani et al., 2013), and the CAESAR-Lisflood (Coulthard et al., 2013) model in a study of Tin Camp Creek in Australia and the Upper Swale in the United Kingdom on a 30 year timescale (Skinner et al., 2018). While these applications use different models, different catchments, different output metrics, and different timescales, there is some overlap in input parameters that allows identification of persistent findings. For example, both studies identify the importance of the sediment transport formula (which we would describe as an alternative model). Additionally, both studies point to the conclusion that the results of sensitivity analysis on landscape evolution models will vary depending on output metrics and the specific study site. For example, while the sediment transport formula is important in both basins examined by Skinner et al. (2018) the relative ranking of other parameters varies within the two basins.

Other applications of MoM within landscape evolution include assessment of the internal dynamics of the LORICA soil profile and landscape model (Temme & Vanwalleghem, 2016), and examination of sediment grain size effects on bedrock river channel morphology using the BRaKE model (Shobe et al., 2018). More extensive reviews of the application of sensitivity analysis methods in environmental and geological models can be found in Pianosi et al. (2016) and Scheidt et al. (2018).

With respect to model calibration, there is a rich history of fitting 1-D model output to observations in order to infer process representation, parameter values, or boundary conditions. Example applications of this approach include fault scarps (Andrews & Bucknam, 1987; Andrews & Hanks, 1985; Hanks, 2000; Pelletier et al., 2006), moraines (Doane et al., 2018), and marine terraces (Rosenbloom & Anderson, 1994). The most extensively examined geomorphic features are river long profiles (Attal et al., 2011; Goren et al., 2014; Fox et al., 2014; Gran et al., 2013; Hobley et al., 2011; Loget et al., 2006; Pelletier, 2007; Pritchard et al., 2009; Roberts & White, 2010; Sweeney & Roering, 2017; Tomkin et al., 2003; van der Beek & Bishop, 2003; Valla et al., 2010).

Extending from 1-D to 1.5-D, Roberts et al. (2012) treats the joint inversion of multiple North American river networks to infer the uplift history of the Colorado Plateau, while Croissant and Braun (2014) inferred detachment-limited stream power parameters based on stream networks extracted from natural and synthetic landscapes. On an event timescale McGuire et al. (2017) tested two alternative debris flow initiation mechanisms.

Calibration of fully two-dimensional planview models has been more limited, and has focused on the production of soil and evolution of hillslopes (Herman & Braun, 2006; Roering, 2008; Pelletier et al., 2011; Petit et al., 2009). Several studies have sought to test and/or calibrate models that combine hillslope and channel processes (Gray et al., 2018; Hancock & Willgoose, 2001; Hancock et al., 2010; Willgoose et al., 2003; Ziliani et al., 2013). The most extensive efforts have focused on the SIBERIA (Willgoose et al., 1991a) and CAESAR-Lisflood (Coulthard et al., 2013) models, which have been calibrated and evaluated on geologic (10,000 year) and subdecadal timescales.

2.5. Multimodel Analysis Strategy

In this work, we follow the guidelines of Hill and Tiedeman (2007) in model development and analysis. Our objective is to identify which model representations of geomorphic process are most consistent with modern topography. We apply the principle of parsimony by starting with a simple model and increasingly adding complexity, such that we can assess when complexity adds value in reproducing modern topography. We use sensitivity analysis to identify the degree to which individual model parameters impact model outputs. To perform these analyses, we used the Dakota software toolkit (Adams et al., 2017a, 2017b) because of its flexible design, availability of a wide range of algorithms for sensitivity analysis and calibration, and support of surrogate-based methods.

Our modeling strategy begins by developing a set of candidate models for the site (section 4). We develop a set of reasonable parameter ranges (section 4.3 and Barnhart et al., 2020c), initial and boundary conditions for our model runs (section 5), and a method to compare models with observed topography (section 6). In this paper we present the results of the sensitivity analysis, which identifies those parameters that most strongly influence the objective function (section 7). The companion manuscript (Barnhart et al., 2020b) presents the calibration and validation efforts. The calibration yields a set of models with parameter values calibrated to minimize model-data misfit. These calibrated models are then validated in a second watershed. Comparing the level of model-data misfit in the calibration and validation results provides a basis for ranking models against one another.

3. Geologic Setting and Identification of Calibration and Validation Watersheds

The study area is located in western New York State, USA, in the Cattaraugus Creek basin, which drains to western Lake Erie (Figure 1). The last glacial retreat from the area left behind a thick accumulation of glacial deposits within the main valleys, including the valleys of the modern Cattaraugus Creek and its tributaries. In the Buttermilk Creek watershed, glacial deposits, together with a thin mantle created by postglacial fan deposits, form a low-relief surface sloping gently downward to the north northwest. Since deglaciation at ~13 ka (Wilson & Young, 2018), Cattaraugus Creek and its tributaries have incised into glacial till deposits (Fakundiny, 1985). Extensive remnants of the postglacial valley surface remain throughout the Buttermilk Creek basin, forming a dissected, semicontinuous, low-relief surface with an altitude that ranges from 400–430 m within the Buttermilk Creek basin. These remnants appear to be only thinly mantled by postglacial deposits (LaFleur, 1979). The low-order channels in the study area are steep-walled ravines and gullies etched into the remnants of a once continuous, low relief surface made up of glacial deposits.

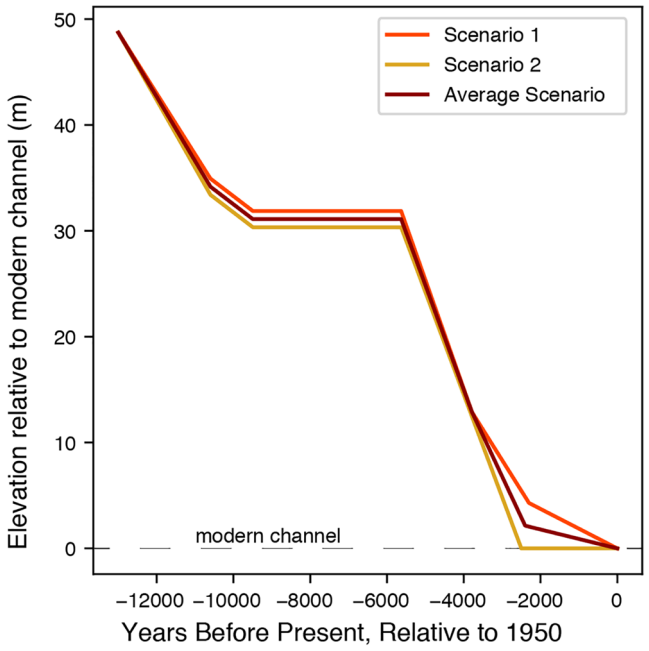

Figure 2. Alternative lowering scenarios for Buttermilk Creek from Wilson and Young (2018).

Geomorphic analysis, ^{14}C dating, and Optically Stimulated Luminescence dating of strath terraces suggest that Buttermilk Creek and its tributaries cut their present-day valleys after the last glacial retreat (Wilson & Young, 2018). The average rate of postglacial downcutting on Buttermilk Creek is ~4 m per thousand years (Figure 2), which is quite rapid by global standards (Milliman & Syvitski, 1992; Portenga & Bierman, 2011; Willenbring et al., 2013). Wilson and Young (2018) identified two alternative scenarios of Buttermilk Creek incision that they determined were equally probable given the available geomorphic and geochronologic data (Scenarios 1 and 2 of Figure 2).

This work focuses on two watersheds in the Cattaraugus Creek drainage. The calibration domain is a watershed located in the Upper Franks Creek watershed in western New York State (Figure 1, also called the "West Valley site"). The validation domain is an unnamed river basin across Buttermilk Creek from the calibration domain. The calibration domain was selected based on the presence of hazardous waste within the watershed boundary. The validation domain was selected because it has similar drainage area, relief, lithology, and base level lowering history to the calibration domain; the study calibration watershed has an area of 4.82 km^2 and total relief of 218 m while the validation watershed has an area of 4.65 km^2 and total relief of 182 m. As the calibration and validation watersheds contain well constrained postglacial surface and incision history of the downstream river, the two watersheds serve as a natural experiment (Tucker & Hancock, 2010).

4. Multimodel Implementation

We identified five categories in which a landscape evolution model of the study site might vary: hillslope processes, hydrologic processes, runoff erosion processes (including channel and gully erosion), site materials, and paleoclimate. We then identified a range of binary options for how the process or property in question might be represented in a long-term erosion model for the West Valley site (Table 1). We selected the 12 binary options based on the following considerations: Does the process or property make sense in the context of the site, the time scale, and the spatial scale? Is there a basis for it in the scientific literature? Is it simple enough, with a sufficiently small number of free parameters, to be practical? Each of the nondefault options in Table 1 represents an *element of complexity* in the model structure. Each change in model structure results in a different governing equation, and thus a different model.

The set of 12 binary options represents $2^{12} = 4,096$ potential alternative model structures. Exploration of this many models is untenable, as each would require an independent sensitivity analysis. This motivated a systematic method for reducing model options. Below we describe our hierarchical approach to model selection, which resulted in a set of 37 alternative models. Each step of model analysis was conducted within this alternative-model framework. All of the computer programs used to implement these models numerically are available through the `terrainbento` python package, which was developed using the Landlab toolkit (Barnhart, Hutton, et al., 2020; Hobley et al., 2017) and is described by Barnhart, Glade, et al. (2019). Landlab is not, itself, a model, but a python package that can be used to construct any number of models. One major benefit of developing alternative models within frameworks such as Landlab is that differences between model governing equations, numerical schemes, and/or other aspects of model implementation can be isolated. Section 4.2 describes the geomorphic basis for each of these options and supporting information Text S1 describes the governing equations.

4.1. Definition of the Hierarchical Model Set

The first model we constructed was the default model, called *Basic* (also known by the model code number 000). The Basic model includes stream power channel incision, uniform lithology, hillslope transport by linear diffusion, and surface water discharge proportional to drainage area. In addition to the Basic model, we constructed 11 single-element models, each of which either adds a new element or alters an existing one. Each of these represents activating the second option in one of the 12 binary options listed in Table 1 (note

Table 1
Binary Options for Process Formulations and Material Properties

Category	Default	Nondefault	Key for Nondefault
Hillslope processes	linear (0)[a]	nonlinear (7)	Ch
Surface-water hydrology	deterministic (0)	stochastic (9)	St
	uniform runoff (0)	variable source area runoff (10)	Vs
Channel/gully erosion	fixed exponent (0)	variable exponent (1)	Vm
	no threshold (0)	erosion threshold (2)	Th
	stream power (0)	shear stress (3)	Ss
	constant threshold (0)	depth-dependent threshold[b](4)	Dd
	detachment-limited (0)	entrainment-deposition (5)	Hy
	uniform sediment (0)	fine vs. coarse[c](6)	Fi
Material properties	no separate soil layer (0)	tracks soil layer (11)	Sa
	homogeneous lithology (0)	two lithologies (12)	Rt
Paleoclimate	constant erodibility (0)	time-varying erodibility[d,e]	Cv

Note. The two-letter key in the rightmost column is used to identify the nondefault option used.
[a]All models are assigned a code and a name. The code is a number calculated using the numbers enclosed in parentheses in the table. The model code for the Basic model is 000. The model code for a model using N_c nondefault choices is calculated as ID = `dec2hex` $\left(\sum_{i=1}^{N_c} 2^{F_i}\right)$ where F_i is the number in parentheses of the ith nondefault choice. The model names are formed by the word Basic followed by the two-letter keys that identify the nondefault option(s) applied. [b]Only applies to models that include a threshold. [c]Only applies to entrainment-deposition models. [d]No factor of (8) present as it was assigned to a model structure with dynamic surface water hydrology that was not computationally feasible to implement. [e]This final permutation was only implemented in one model, and the three-digit model ID was assigned manually to CCC.

that one of these—the addition of a factor to represent the fraction of fine-grained sediment—was omitted from the single-element model set because it is inapplicable as a stand-alone option). Comparison of each of these models with model Basic serves as a test of the ability of the new element of complexity to simulate the modeled system.

Assuming all binary options are independent and can be combined, twelve binary options results in 66 two-element models. Among this potential set, we omitted several combinations that were logically incompatible (for example, it would not make sense to combine a depth-dependent threshold option (Dd) with a fixed-threshold option (Th)).

The two-element models have two purposes:

1. To identify, for a given successful one-element model, the additional elements of complexity that further improve simulation performance.
2. To identify nonlinear interactions in model structure.

Starting with all possible two-element models, we identified models that were reasonable to omit from our analysis. We retained all models for which we could identify a plausible mechanism for nonlinear interaction between the included components. Combining stochastic precipitation (St) and fluvial erosion threshold (Th) is an example of such a two-element model combination, because the threshold dictates what subset of stochastic events will produce geomorphic change at a given location (e.g., Tucker & Bras, 2000). We omitted models for which we could not identify such a mechanism. An example of a combination without an obvious mechanism for interaction would be the use of a nonlinear hillslope rule (Ch) together with a water-erosion law that treats the drainage area exponent as a free parameter (Vm). When we were uncertain regarding the presence of a plausible nonlinear interaction, we erred on the side of caution and retained the combination.

The process of model identification and construction resulted in a total of 36 distinct models. Based on calibration results, we also constructed one three-element model, for a final total of 37 alternative models (Table 2).

Table 2
Summary of Individual Models

Model code and name	`terrainbento` program if different	Element varied #1	Element varied #2	Element varied #3
000 Basic	Basic	—	—	—
001 BasicVm	Basic	variable m	—	—
002 BasicTh		threshold	—	—
004 BasicSs	Basic	shear stress[a]	—	—
008 BasicDd		$\omega_{ct} \propto$ incision depth	—	—
010 Basic Hy		entrainment-deposition[b]	—	—
040 BasicCh		nonlinear creep	—	—
100 BasicSt		stochastic runoff	—	—
200 BasicVs		VSA[c]	—	—
400 BasicSa		tracks soil/alluvium	—	—
800 BasicRt		tracks two lithologies	—	—
CCC BasicCv		K varies over time	—	—
012 BasicHyTh	BasicHy	variable ω_c	entrainment-deposition	—
102 BasicStTh		variable ω_c	stochastic runoff	—
202 BasicThVs		variable ω_c	VSA	—
802 BasicRtTh		variable ω_c	tracks two lithologies	—
00C BasicDdSs	BasicDd	shear stress	$\omega_{ct} \propto$ incision depth	—
014 BasicHySs	BasicHy	shear stress	entrainment-deposition	—
104 BasicSsSt	BasicSt	shear stress	stochastic runoff	—
204 BasicSsVs	BasicVs	shear stress	VSA	—
804 BasicRtSs	BasicRt	shear stress	tracks two lithologies	—
018 BasicDdHy		$\omega_{ct} \propto$ incision depth	entrainment-deposition	—
108 BasicDdSt		$\omega_{ct} \propto$ incision depth	stochastic runoff	—
208 BasicDdVs		$\omega_{ct} \propto$ incision depth	VSA	—
808 BasicDdRt		$\omega_{ct} \propto$ incision depth	tracks two lithologies	—
030 BasicHyFi	BasicHy	entrainment-deposition	variable fraction fines	—
110 Basic HySt		entrainment-deposition	stochastic runoff	—
210 BasicHyVs		entrainment-deposition	VSA	—
410 BasicHySa		entrainment-deposition	tracks soil/alluvium	—
810 BasicHyRt		entrainment-deposition	tracks two lithologies	—
440 BasicChSa		nonlinear creep	tracks soil/alluvium	—
840 BasicChrRt		nonlinear creep	tracks two lithologies	—
300 BasicStVs		stochastic runoff	VSA	—
600 BasicSaVs		VSA	tracks soil/alluvium	—
A00 BasicRtVs		VSA	tracks two lithologies	—
C00 BasicRtSa		tracks soil/alluvium	tracks two lithologies	—
842 BasicChRtTh		nonlinear creep	tracks two lithologies	threshold

[a] Shear stress version of water erosion term. [b]Entrainment-deposition ("hybrid") water erosion law. [c]Variable source area hydrology.

The hierarchical approach to model construction has several advantages. We can compare calibrated versions of each one-element model to identify (a) how each element alters model behavior, and (b) whether any of those process elements improve performance relative to the simplest model. Then, for those elements that do improve performance, we can select their corresponding two-element models and repeat the comparison. In this way we systematically assess the context in which model complexity improves (or fails to improve) simulation performance. We assigned each model a three digit hexadecimal ID code and a name. This code was determined by giving each of the twelve choices a factor number (shown in parentheses in

Table 1). Each of the nondefault options was given a two-letter key such that each model has a unique name. When a model uses multiple options, the two letter keys are added in alphabetical order. For example, the simplest model has the code 000 and is called Basic, whereas a model that adds the ability to track a regolith layer has the code 400 and is called BasicSa.

4.2. Basis for Geomorphic Process Options

The geomorphic process options presented in Table 1 were selected from existing theory based on site observations. Next we review the basis for each of these binary choices. Text S1 presents the mathematical form of the governing equations and Barnhart, Glade, et al. (2019) provides a full derivation of each model's governing equations.

4.2.1. Hillslope Processes

Prior work at the West Valley site (Wilson & Young, 2018) and our own observations support the conclusion that geologic material is transported downslope by two primary processes: landsliding and soil creep. Linear and nonlinear diffusion theories have both been successfully used to model slope forms in cases where soil creep—downslope motion of soil that often arises from repeated disturbance and displacement of material by processes such as bioturbation—is considered to be the primary source of transport (e.g., Culling, 1963; Gilbert, 1909; Roering, 2008). Given the above considerations, we identify two alternative options for modeling downslope soil transport: linear creep theory (default), and nonlinear creep theory.

Unfortunately, we presently lack a universally agreed upon transport law for landsliding *sensu stricto* (see, for example, Dietrich et al., 2003). That said, linear diffusion theory has been used as a rough proxy for the long-term effects of landsliding (e.g., Willgoose et al., 1991a, 1991b), though it does not account well for the expected acceleration as a hillslope approaches an effective threshold angle for soil stability. Nonlinear diffusion theory has also been used as a proxy for the integrated effects of shallow landsliding, and has the advantage of capturing accelerated soil movement as the slope angle rises toward an effective angle of repose (Andrews & Bucknam, 1987; Roering, 2008). To our knowledge, no effective process law has been demonstrated for long-term erosion by deep-seated rotational failures. Early work on incorporating landslides in landscape evolution models considered failure based on a slope stability criterion (Densmore et al., 1998). More recent work suggests that earthflows—elongated and fairly deep forms of mass wasting that behave somewhat like "glaciers of mud"—may be effectively modeled using an approach that shares some important elements with nonlinear diffusion theory (Booth & Roering, 2011; Booth et al., 2013).

4.2.2. Hydrologic Processes

It is not computationally feasible to represent the full hydrodynamics of precipitation, infiltration, and runoff generation on geologic timescales. A simple and common approach, which serves as our default option, is to use drainage area as a surrogate for water discharge, and thus erosion by streamflow (e.g., Willgoose et al., 1991a, their Appendix B). This approach implies that information about precipitation is embedded in any parameters that control the effectiveness of erosion by flowing water.

An alternative is to use a stochastic approach, treating river discharge or precipitation rate as a random variable with a defined probability distribution (Deal et al., 2018; DiBiase & Whipple, 2011; Lague et al., 2005; Snyder et al., 2003; Tucker, 2004; Tucker & Bras, 2000). Use of a stochastic approach has the advantage of capturing a spectrum of natural events. The primary disadvantage is complexity: the stochastic approach tested in this study requires specification of four additional parameters.

Both hydrology models described above use the assumption that the average runoff rate is spatially uniform. Yet as early as the 1970s, hydrologists discovered that certain areas of a drainage basin may sometimes contribute substantially more runoff than others. In vegetated, humid areas such as western New York, runoff can be produced by saturation-excess rainfall (Dunne, 1970; Dunne & Black, 1970). The propensity of the land to become saturated, and thus produce runoff, is concentrated in areas with low slopes or high drainage area (Beven & Kirkby, 1979; Dietrich et al., 1993; Dunne & Black, 1970; O'Loughlin, 1986). This description of hydrology is called variable source area (VSA) and is provided as an additional option. The mathematical representation of variable source area hydrology derives from work by O'Loughlin (1986) and Dietrich et al. (1993), and is described in Text S1.4.

4.2.3. Erosion by Channelized Flow

The geomorphology community has devoted considerable effort toward understanding the physics of channel incision into cohesive and/or rocky material (see, e.g., review by Whipple, 2004). Despite this effort, the

community is still debating the core principles governing channel incision, and how those principles depend on considerations such as material properties and the time scale of interest.

As a starting point, it is useful to identify those elements of contemporary channel incision theory that do *not* apply especially well to the West Valley site. Among these options are the saltation-abrasion theory (Chatanantavet & Parker, 2009; Sklar & Dietrich, 2004), which considers wear of bedrock to occur primarily by abrasion by saltating grains. At West Valley, most geologic materials subject to erosion either: (1) are cohesive but mechanically weak sediment (such as the clay-rich Lavery Till), or (2) possess a high density of fractures and/or bedding planes (as in the shale-rich sedimentary units). Moreover, we have not observed classic abrasion-related features such as flutes and potholes in the study site's channels. We conclude that abrasion is likely to be a minimal contributor to the erosion of these materials as compared with direct hydraulic detachment. The site's geologic materials clearly possess high cohesive strength (sufficient, for example, to maintain near-vertical steps and waterfalls), which suggests that a conventional transport-limited model would also be inapplicable (for a general discussion and comparison of these types of models, see for example Whipple & Tucker, 2002).

Two general classes of channel erosion model remain for consideration. The first is known as *detachment-limited* theory, because it assumes that the primary limitation on the rate of channel downcutting is the rate at which particles can be detached rather than the rate at which they can be transported (Howard, 1994). There exist a variety of different forms of the detachment-limited model. The most common forms of this model assume that the rate of channel downcutting depends on either hydraulic power (*stream power* per unit channel width) or friction force (*boundary shear stress*). Usually power or force is represented as a power function of channel slope, water discharge (or drainage area as a surrogate), and a factor that lumps together information about material properties, roughness, and channel dimensions (e.g., Howard & Kerby, 1983; Whipple & Tucker, 1999). Sometimes an erosion threshold is included in the formulation such that when power or force falls below this threshold, little or no incision occurs (e.g., Lague et al., 2005; Tucker, 2004).

The other major class of channel erosion law adds sediment mass conservation to detachment-limited theory by allowing for the possibility of sediment deposition. There is no general name for such models; here we refer to them as *entrainment-deposition* models because they include terms for both entrainment of material from the channel bed into the flow, and deposition from the flow onto the bed. The essence of this approach is that the net rate of erosion is equal to the difference between the rate of particle detachment from the bed and the rate of deposition from active transport. Entrainment-deposition models have been used to simulate landscape-scale erosion (e.g., Davy & Lague, 2009; Shobe et al., 2017) and process-scale gully erosion (Rengers & Tucker, 2014; Sidorchuk, 1999). In tracking sediment detached from the bed, one has the option of distinguishing between coarser material that might be redeposited, and fine (silt and clay) materials that are effectively removed from the channel network as soon as they are detached. Such an approach was suggested by Kirkby and Bull (2000) and recently implemented by Shobe et al. (2017). This type of model permits exploration of the "cover" effect of the saltation-abrasion models (Sklar & Dietrich, 2004).

This brief review of channel incision theory suggests at least two candidate process laws for fluvial incision. Detachment-limited models may be appropriate at our field site given the cohesive but fine-grained nature of the geologic materials. On the other hand, entrainment-deposition models allow for the possibility of sedimentation in some locations. However, considering just these two models entails a number of other binary choices that may dictate model success including: whether to use a fixed or variable exponent on the discharge factor; whether to include an erosion threshold; whether to use a shear stress or stream power formulation of erosion potential; and whether to distinguish between fine and coarse material (in the entrainment-deposition models only). Furthermore, studies of rapid postglacial erosion in the Le Seuer River basin, Minnesota provide one additional option to consider (Gran et al., 2013). Based on observations of downstream coarsening, they describe a model in which the erosion threshold increases with progressive incision. Given the similarity between the Le Seuer case and the West Valley Site, we included an option to allow the erosion threshold to increase with progressive incision depth. Recent work by Bennett (2017), however, indicates that channels surveyed near the West Valley Site do not exhibit downstream coarsening, but instead have relatively uniform bed material size.

4.2.4. Representation of Geological Materials

We considered two choices associated with representation of geologic materials. The first choice pertains to the site's geologic units and takes as a default homogeneous erodibility of all materials. This option is based on the hypothesis that shale and clay-rich till are similar when it comes to erosion by running water or gravitational processes; both contain large fractions of clay size material, both are mechanically weak, yet both possess enough cohesion to support vertical faces. The alternative is to distinguish between glacial deposits (hereafter sometimes referred to as "till" for simplicity) and bedrock. This second approach represents the hypothesis that bedrock is significantly more resistant to erosion than the till. Including both options in the modeling framework tests the degree to which incorporating a distinction between rock and glacial sediments increases an erosion model's performance. We considered subdividing each of the two units further, but chose not to so as to strike a balance between fidelity to the differences in erosional susceptibility among different units and the need to keep models simple enough to calibrate.

An additional choice concerns the representation of a layer of mobile regolith as distinct from underlying parent material. The default approach is to treat the soil-rock continuum as being effectively homogeneous. Incorporating a dynamic regolith layer has the appeal of realism, but it requires a method to calculate the creation of new regolith from the underlying parent material (e.g., Ahnert, 1976; Heimsath et al., 1997; Small et al., 1999), as well as a rule to ensure that hillslope transport does not exceed the available regolith supply. The West Valley Site's region is predominantly soil-mantled, but apart from that observation we have no a priori reason to prefer one approach over the other. We have therefore included, as one of the binary choices, the option to include or not include a dynamic regolith layer.

4.2.5. Paleoclimate

The final model structure variation we considered stems from uncertainty about variations in paleoclimate over the 13 ka calibration period. We can gain some insight into the region's postglacial climate history from the TraCE-21ka model experiment and data set, which derives from a continuous global simulation of climate evolution over the last 21 ka (Liu et al., 2009, http://www.cgd.ucar.edu/ccr/TraCE/). For the model grid cell that covers western New York State, the modeled precipitation amounts stay roughly constant over the past ~13 ka but change in character (Figure S1). The biggest change is in convective precipitation, which rises from ~13 to ~6 ka then stays roughly constant. Large-scale stable snow is modeled as decreasing steadily over the last ~13 ka, with the trend flattening out by about ~4 ka.

To assess the degree to which incorporating the effects of variable paleoclimate improves model performance, we developed an additional model that incorporates the possibility of past variations in the stream power erodibility coefficient K, which embeds the properties of climate that alter the characteristic discharge. In this model variant, a fixed value is selected for K for the time period from 5 ka to the present day. A factor f between 0.5 and 1.5 is used such that the stream power erodibility at the beginning of the model run (representing 13 ka) is fK. The $\pm$50% range in f equates to varying the mean daily rainfall from 25–225% of its final (present-day) value, based on a model that links K with precipitation parameters (as described in Barnhart, Glade, et al. (2019, section 4.2)). Lacking constraints, we chose a large range.

4.3. Model Parameters

Each of the 37 alternate models takes between two and ten input parameters (Table S2). For each of these parameters we identify a range of values considered reasonable given the current literature and the location of the study site (see Text S5 and Barnhart et al., 2020c for details). Table S1 summarizes the range used in sensitivity analysis and calibration.

4.4. Space and Time Discretization

Discretization in space and time was designed to balance process representation with the goal of keeping model run times close to one half hour. We use a global model time step of 10 years (note that some individual process components will subdivide this global step as needed to ensure numerical stability; Barnhart, Glade, et al., 2019). A notable exception is the stochastic precipitation models (St), which use between 1 and 20 sub–time steps.

In both the calibration and validation watersheds, we used a horizontal grid cell spacing of 7.3 m (24 ft), which represents a resolution somewhat finer than the 10 m spacing recommended by Zhang and Montgomery (1994) as an appropriate compromise between resolution and data volume for flow-routing applications. With this grid cell size, the smallest hillslopes are represented by at least 10 grid cells. With

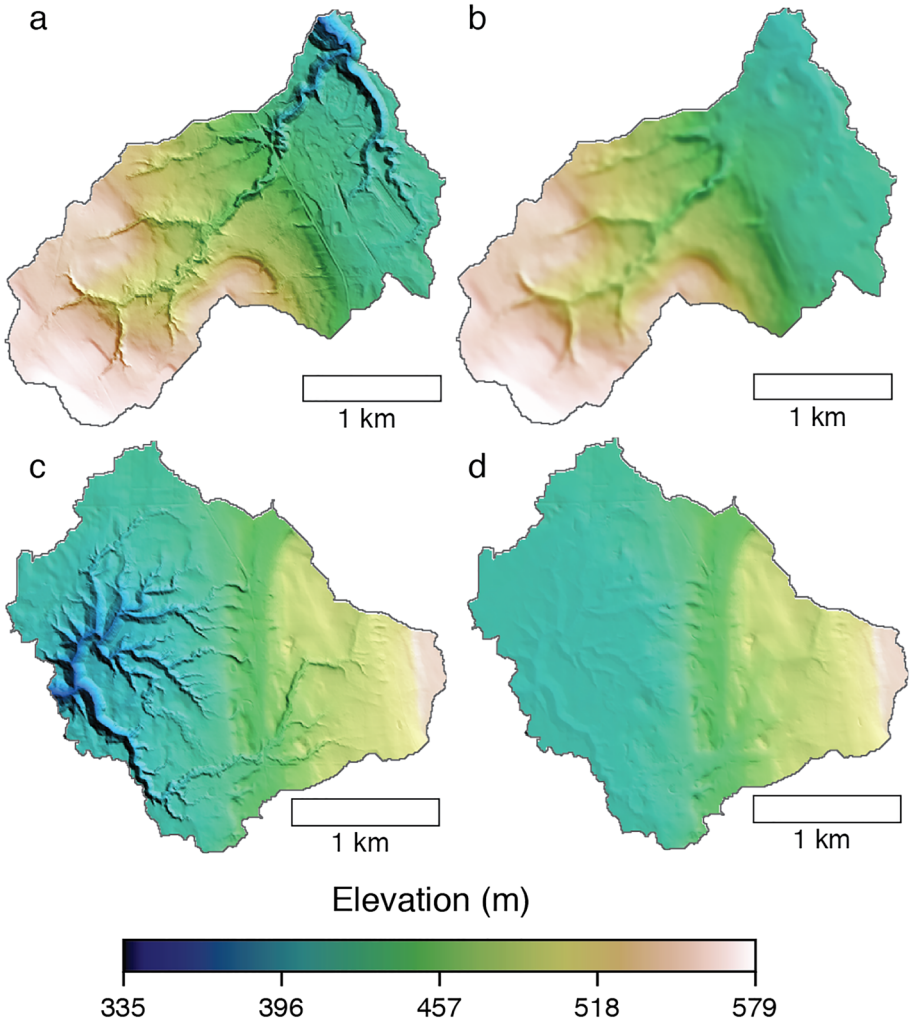

Figure 3. Shaded relief maps of modern (a, c) and postglacial (b, d) topography for the calibration watershed (a, b) and validation watershed (c, d). The 0% etch condition is shown here. Figures S3 and S4 show equivalent maps of all initial condition grids.

a value of $K = 0.001$ yr^{-1}, this grid cell size corresponds to a Courant-condition stability criteria of 0.3 at the watershed outlet for the advection portion of the Basic model governing equation (equation (S5)). The Courant-condition value is given as the ratio of time step size to the erosion wave traveltime for one cell. Note that the units of K vary based the value of the drainage area exponent m.

5. Initial and Boundary Conditions

Model simulations require a starting topographic surface (digital elevation model, or DEM) and information about the boundary conditions to the study watershed through time. Models that differentiate between till and bedrock also require a starting depth-to-bedrock.

The time-variable rate of downcutting on Buttermilk Creek (Figure 2) serves as the model domain boundary condition. We considered both lowering scenarios in the sensitivity analysis and used the average for calibration. The depth to bedrock varies from surface exposure to a few tens of meters (Figure 1). Depths are based on Randall (1980, see Figure S5).

The starting topographic surface represents the preincision valley topography as it existed following the initial retreat of the ice sheet. As the remnants of the plateau described in section 3 appear to be only thinly mantled by postglacial deposits, it is logical to assume that they provide a reasonably accurate representation

of the postglacial valley topography shortly before stream incision began. Using plateau remnants for guidance, we constructed the preincision valley topography by modifying the present-day topography (Figure 3). Starting with a LiDAR data set for the site (Cortes, 2016), we constructed a gridded data set at 7.3 m (24 ft) and contoured it. Areas identified as plateau remnants were not altered. Where stream channels dissected the postglacial surface, we edited the contour lines to extrapolate the remnant plateau across the incised areas (contour lines shown in Figure S2). Adjusted contours were smoothed and converted to an elevation grid using the topo-to-raster interpolation technique (ESRI, 2014). This technique yielded preincision surfaces with a gently sloping valley that preserves the overall slope observed in plateau remnants.

We then used the preincision valley topography as the basis for constructing a set of six alternative initial condition DEMs (Figures S3 and S4). This set of alternative initial conditions was used in sensitivity analysis to test three sources of elevation uncertainty.

First we created four permutations that lightly etched the present-day drainage network into the preincision topographic surface. This etching procedure, which has been used in other landscape modeling studies (e.g., Anderson, 1994), does not substantially alter the macroscopic erosion patterns. Etching helps increase the number of solutions in which computed patterns of erosion depths are comparable to the present day by reducing the number of simulations where the main streams are shifted to one side or the other in the main valley due to local discrepancies in the location of the channels. We etched the preincision valley topography surface by 0%, 3.5%, 7%, and 14% of the difference between the present-day and preincision valley topographic surfaces.

We also constructed two additional input DEMs: one in which normally distributed random noise of 0.04 m (0.132 ft, the vertical accuracy of the LiDAR data set) was added to the 7% etched DEM. This input DEM was intended to identify the effect of the LiDAR observational uncertainty on model outputs. A final DEM used the 7% etching case but left the upper portions of the watershed unchanged.

6. Model-Data Comparison

In order to perform a sensitivity analysis and calibration of each model, we need a basis to compare observed and simulated topography and evaluate model fit. At our site, our only observable is the modern topography. As discussed in section 2.3 there is no established approach to compare end-of-simulation and modern topography. We assessed sensitivity based both on a set of statistical topographic metrics, and on a direct pixel-by-pixel comparison between end-of-simulation topography and modern topography. Model-data comparison employed the `umami` python package (Barnhart, Hutton, et al., 2019).

As has been done in several prior studies, we evaluate model sensitivity using a set of statistical metrics derived from the DEM. These metrics included ones that have been previously used in comparisons of observed and modeled topography, such as the hypsometric curve, cumulative area distribution, and the statistical moments of the elevation and slope distributions (e.g., Hancock & Willgoose, 2001; Ibbitt et al., 1999; Hancock et al., 2010, 2011; Howard & Tierney, 2012; Perera & Willgoose, 1998; Skinner et al., 2018). Many of the most common approaches to evaluating topography do not aggregate topographic data into a single value (for example, the hypsometric curve is a frequency distribution rather than a scalar value). To capture the shape of the cumulative area distribution, and in particular the small-area and large-area ends of the distribution, we use the number of nodes with drainage area of one, two, three, and four model grid cells, and the 95th, 96th, 97th, 98th, and 99th percentiles. We evaluated the cumulative distribution of elevation for major inflection points, and selected the 2nd, 8th, 23rd, 30th, 36th, 50th, 75th, 85th, 90th, 96th, and 100th percentiles.

We also developed a direct cell-by-cell comparison of modeled and observed DEMs. The direct comparison approach is feasible for the West Valley site because remnants of the postglacial, preincision topography are well preserved and we expect that model stream channels will incise in the same locations as modern stream channels. While we are uncertain of the minor details of the paleotopography (as discussed in section 5), this site does not suffer from uncertainty in the macroscale paleotopography.

Our intention is to reduce the topographic difference to a lower dimension than the number of grid cells in the model DEM (e.g., 10^5 for this study) while grouping grid cells into domains with similar dominant process (e.g., channel bottoms, valley side slopes, till plateau areas, hillslope crests). In the calibration presented by Barnhart et al. (2020a) we use a nonlinear least squares optimization approach to identify parameter

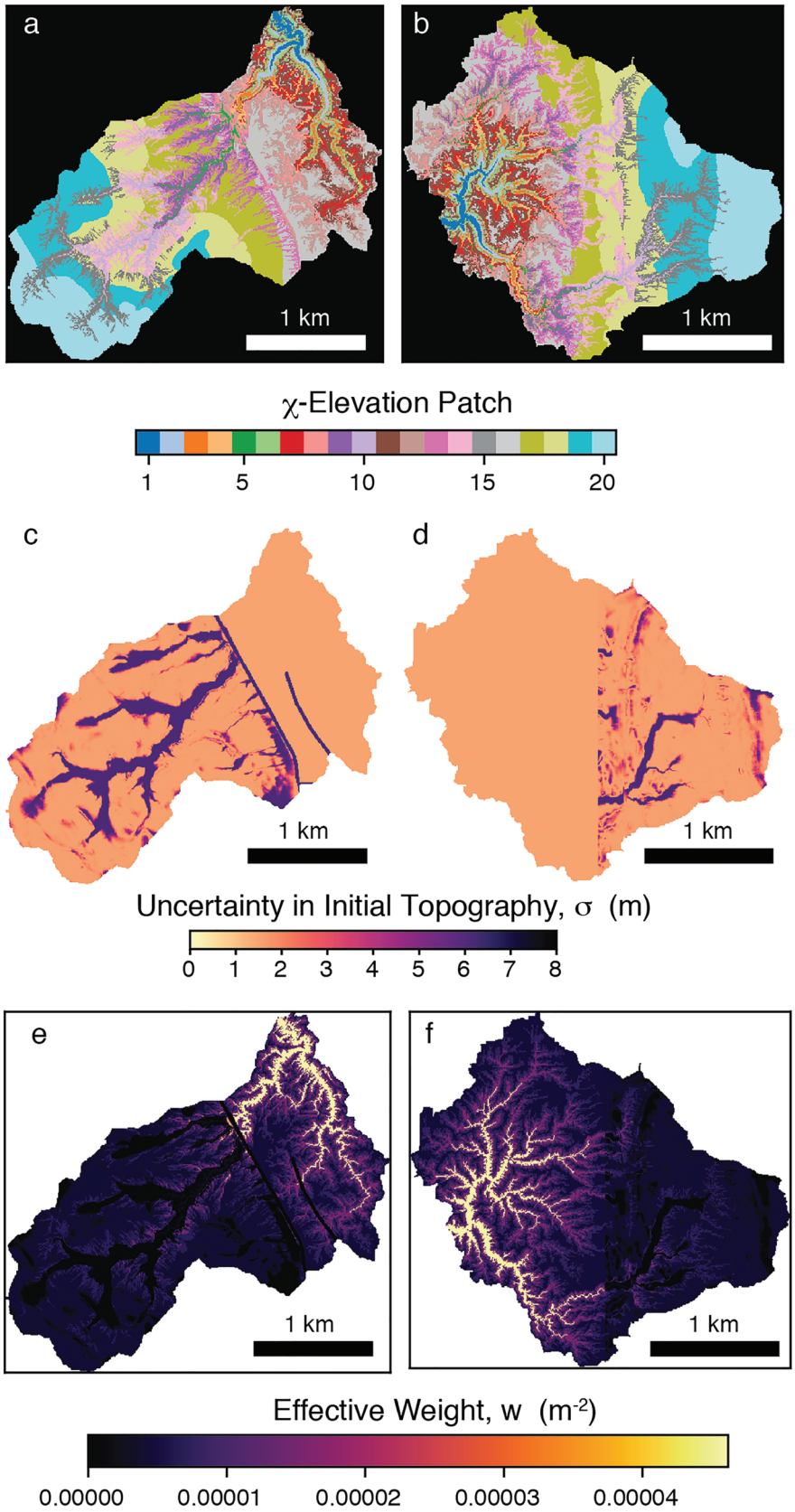

Figure 4. Maps of χ elevation categories (a, b), initial condition uncertainty (c, d), and effective weights w_i (e, f) for the calibration domain (a, c, e) and the validation domain (b, d, f).

values that minimize the difference between model simulation results and modern topographic data. This method uses a *sensitivity matrix* (also called the Jacobian), which describes the partial derivatives of each simulated equivalent with respect to each calibrated parameter.

To reduce our 10^5 data points into a computationally feasible number, we divided the landscape into a set of 20 patches that broadly represent characteristic landform features, such as canyons cut into till or bedrock uplands. Grid cells in the DEM were divided into categories based on two criteria: elevation, and the χ index value (Perron & Royden, 2013). χ is a property of channel networks, defined as the upstream integral of a power function of drainage area evaluated from base level to a given position, that can be related to the expected equilibrium elevation field of an erosional channel network (Harkins et al., 2007). As such, it is a useful frame of reference with which to identify process domains that may be spatially disparate, but share similar local base level histories. Taken together, we found that these two topographic properties broadly characterized the landscape into distinct landform features, and provided a semiautomated method to identify comparable patches in both the calibration and validation sites with which to construct equivalent metrics of model success (Figures 4a and 4b). We divided the catchment DEM into five χ index categories based on percentile of the χ value (0–5, 5–20, 20–50, 50–100). For each of these χ domains, we divided the domain into four elevation bands, again based on percentile (0–25, 25–50, 50–75, 75–100). This approach is similar to splitting the catchment into multiple subbasins based on stream order, as is done in Skinner et al. (2018).

For each patch j, a model misfit score P_j is calculated as the square root of the sum of squares of the residual between observed and modeled elevations in the patch's grid cells

$$P_j = \sqrt{\sum_{i=1}^{N_j} w_i \left(\eta_i^{\text{obs}} - \eta_i^{\text{sim}}\right)^2}, \tag{2}$$

where η_i^{obs} is the observed elevation at cell i, η_i^{sim} is the simulated equivalent, N_j is the number of grid cells in patch j, and w_i is a cell level residual weighting factor defined below (equation (3)). The fundamental element of the objective function is the square of each model grid cell's residual (the L_2 norm loss function).

To calculate weight factors for individual grid cells, an initial-condition uncertainty, σ_i, is combined with the total number of grid cells in a particular patch. The weight factor for cell i which belongs in patch j is

$$w_i = \frac{1}{\sigma_i^2 N_j}, \tag{3}$$

where N_j is the total number of grid cells in patch j. Including the number of cells in the patch provides a method for weighting some patches more than others. Patches with larger patch-averaged σ^2 and larger N_j will be less influential in model fit. This allows us to identify certain landscape features as having more importance in the model calibration. For example, patches that contain the lower reaches of Franks Creek, Quarry Creek, and Erdman Brook are considered to be especially important because they reflect rapid postglacial downcutting in the glacial materials that underlie the study watershed. Additionally the valley side slopes adjacent to these channels and the till plateau surface are given more importance than the upper portion of the watershed. The net result are effective weights that emphasize the incised channels and gullies within the till plateau area (Figures 4e and 4f).

The objective function is defined simply as the sum of the squared patch scores:

$$F_{\text{obj}} = \sum_{j=1}^{M} P_j^2. \tag{4}$$

The objective function therefore includes the combined misfits of all 20 patches. The patch scores, P_j^2, themselves are weighted equally, but their individual grid cells take on different weights, which depend on the size of the patch and an estimate of the uncertainty associated with postglacial erosion at a particular cell, as discussed below.

Figures 4c and 4d present our estimates of initial-condition uncertainty, σ [L], assigned to the grid cells of each watershed. In the plateau areas of the landscape, geologic evidence suggests that little erosion has occurred (Wilson & Young, 2018).

Along the lower reaches of streams that cut the plateau surface (such as Franks Creek), the total depth of postglacial erosion is constrained because the plateau serves as a reference surface. Thus, the initial-condition uncertainty associated with these locations is considered to be relatively low. In this area, and in the bedrock hillslopes in the upper watershed, $\sigma = 1.52$ m (5 ft, orange color in Figures 4c and 4d). In the upper parts of the drainage network, above the glacial till plateaus, we lack a reference surface. The incised valleys in these locations might have been carved since the last glacial retreat or they might have been carved earlier. Because we do not know their incision history, these locations have greater initial-condition uncertainty. Based on the approximate depth of channels in the upper watershed relative to the surrounding hillslopes, we set $\sigma = 6.01$ m (20 ft) in these areas (purple color in Figures 4c and 4d). We also assign the larger value of $\sigma = 6.01$ m to locations of strong anthropogenic modification within the modeled domains, including roads.

7. Sensitivity Analysis

Sensitivity analysis is a formalized method for documenting a model's dynamics by systematically varying model inputs and measuring the impact of those variations on model output. In addition, our application focuses on screening of initial conditions, boundary conditions, and parameter estimates so that we can reduce the parameter space considered in the calibration (Barnhart et al., 2020b). Specifically, the goals of our sensitivity analysis are to:

1. Explore the simulated dynamics of the system.
2. Assess the degree to which alternative reconstructions of postglacial topography influence simulated equivalents.
3. Evaluate the degree to which plausible base level lowering histories influence simulated equivalents.
4. Identify which parameters need to be included in calibration and which parameters are insensitive, and thus can be held constant during calibration.

These goals are similar to other applications of sensitivity analysis in landscape modeling (Shobe et al., 2018; Skinner et al., 2018; Temme & Vanwalleghem, 2016; Ziliani et al., 2013).

7.1. Experimental Design

We applied sensitivity analysis to the calibration watershed and considered a full factorial design across the 36 zero-, one-, or two-element models using two alternative base level lowering histories (Figure 2) and six alternative initial conditions (section 5). We used both the topographic metrics and the objective function defined in section 6.

Because our models each require between two and ten parameters, and individual model evaluations take on order ~30 min, we are not able to assess sensitivity through grid search and line search approaches, which scale geometrically with the number of parameters. Still, we seek a measure of the global sensitivity of the simulated equivalents to the model parameters. We used a Method of Morris (MoM) screening method (Campolongo et al., 2007; Morris, 1991) because our parameters are continuous variables, MoM handles nonlinearities well, and the number of required model evaluations scales linearly with the number of model parameters.

After all model runs are complete, MoM calculates two global statistics for each combination of input parameter and simulated equivalent. The first statistic is μ^*,the mean of the absolute value of the elementary effects, which measures the overall parameter importance. Using the absolute value reduces the likelihood of Type II errors (Campolongo et al., 2007). The second statistics is σ^*, the standard deviation of the absolute value of the elementary effects, which provides a global measure of variability of elementary effects for that parameter (see Saltelli et al., 2008, p. 117).

The MoM was designed to screen parameters into three primary categories: those parameters with effects on an outcome that are (a) negligible such that μ^* and σ^* are both small, (b) linear and additive such that μ^* is large but σ^* is small, and (c) nonlinear such that both μ^* and σ^* are large.

This method is well suited for applications with high computational cost (Saltelli & Annoni, 2010) and has been applied to other surface processes models (e.g., Shobe et al., 2018; Skinner et al., 2018; Temme & Vanwalleghem, 2016). Figures 5–7 show sensitivity in a selection of models and Table S2 lists the parameters

varied for each model. For parameters that plausibly span many orders of magnitude, we used the $\log_{10}$ of the parameter for both sensitivity analysis and calibration.

The computational load required for this analysis was significant. Even with a computationally frugal method such as MoM, which scales linearly with the number of model parameters, the described sensitivity analysis required ~30,000 model evaluations, with an average run time of ~30 min each.

7.2. Method of Morris

MoM is a global sensitivity analysis method that considers a k-dimensional space associated with k model input parameters. The extent of parameter space is set by user-defined parameter ranges (see Table S1 for ranges used in this study). Samples are taken from across the entire parameter space, so the results represent global sensitivities. MoM makes a series of trajectories through parameter space by starting from randomly chosen points. The method changes the value of one parameter at a time by an amount around half of that parameter's range. Importantly, this characteristic of MoM means that its results depend on the parameter ranges. In each trajectory, the order in which parameters change is different. For each parameter and each simulated equivalent, the trajectory defines one *elementary effect*, a measure of how changing the input parameter changes the simulated equivalent. Using at least four to ten trajectories is recommended (Morris, 1991; Saltelli et al., 2004), and we use ten. μ^* and σ^* are calculated using the elementary effects.

7.3. Assessing Sensitivity to Categorical Inputs

To assess the sensitivity of model outputs to the outlet downcutting trajectory and initial condition, we ran the same parameter set sequences within each combination of downcutting history and initial condition. This was accomplished by using the same random seed value for all initial and downcutting combinations for a given model. This allowed us to calculate sets of (μ^*, σ^*) values that measure the sensitivity of model outputs to the downcutting history and initial condition. The calculation of these elementary effects used the "7% etch" initial condition and Scenario 1 as the reference case. This approach to measuring the sensitivity of a categorical input is similar to the method used by Skinner et al. (2018) to test sensitivity to the choice of sediment transport formula.

8. Results

We focus on assessing (1) the model sensitivity to initial and boundary conditions, and (2) identifying which parameters are most and least important for model outputs.

8.1. Topographic Metric Sensitivity

Figure 5 presents a summary of the sensitivity analysis results for one of the 36 models, based on the 25 statistical topographic metrics that we considered. We focus here on just one of the 36 models, because its sensitivities in terms of the terrain metrics are illustrative of the kind of sensitivity we observed across the full suite of models. Results for model 802 are shown as it is the best performing two-element model in the calibration (Barnhart et al., 2020b). This model distinguishes between till and rock, and includes erosion threshold parameters for both lithologies (ω_{c1} and ω_{c2} for till and rock, respectively). The four parameters that control channel erosion consistently have the highest mean sensitivities. Unsurprisingly, the parameters that control channel erosion into till (K_1 and ω_{c1}) dominate the lower elevation percentiles. The maximum elevation is most sensitive to the erosion thresholds because thresholds can shut off erosion at the highest elevation portions of the landscape that have low drainage area and slope.

In general the hillslope diffusivity D, the initial condition, and the lowering history exert little influence on this set of topographic metrics. Notable exceptions include sensitivity of the high percentiles of the cumulative area distribution to D and the initial condition. These high percentile values are influenced by the details of channel branching in the lower part of the watershed.

8.2. Sensitivity of Objective Function

8.2.1. Initial and Boundary Conditions

The objective function (equation (4)) shows little sensitivity to the difference between the two lowering history scenarios shown in Figure 2. Model 802 BasicRtTh, for example, shows much greater sensitivity to model parameters than to the choice of initial topography or lowering history (Figure 6b).

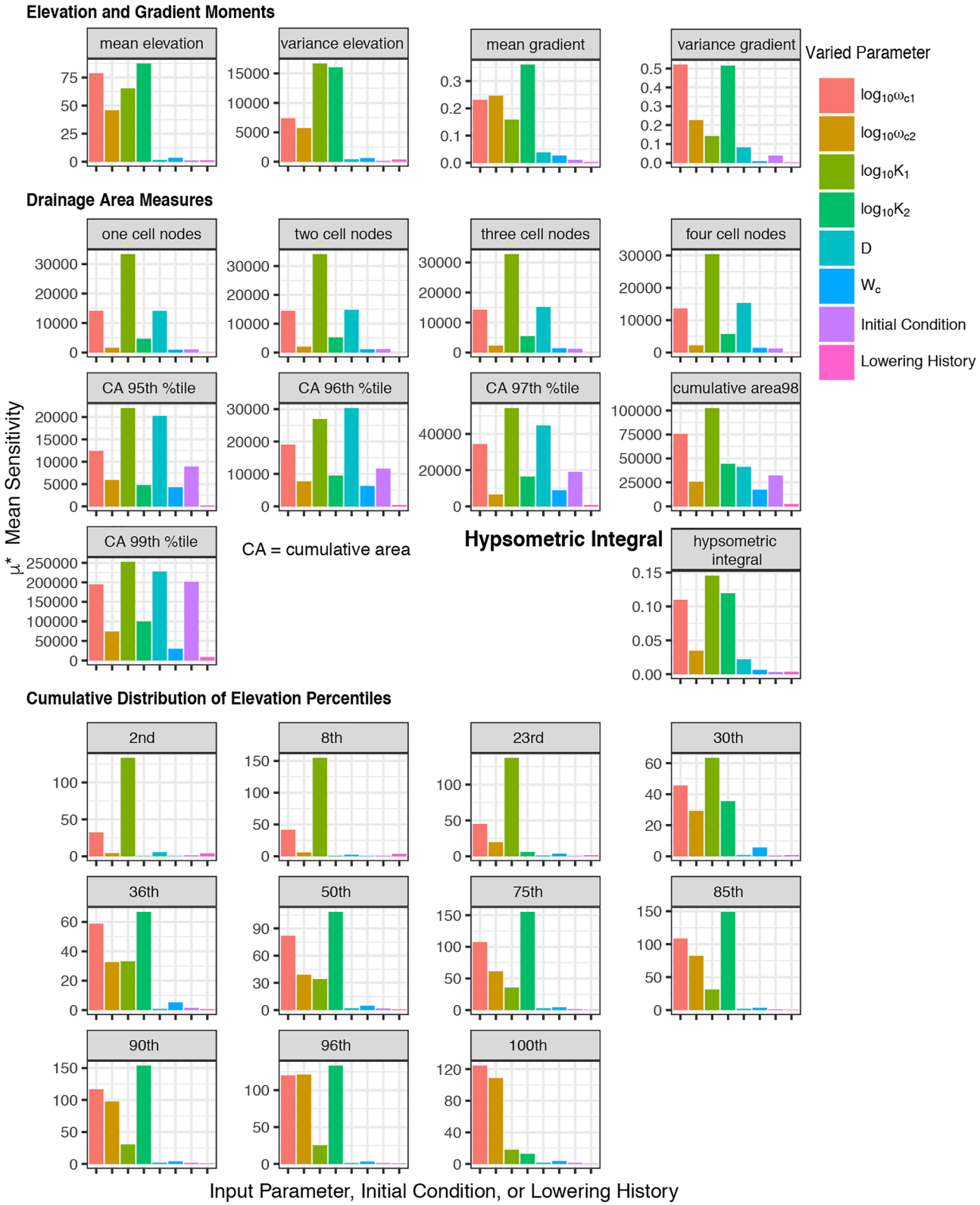

Figure 5. Values of μ^* for the 25 topographic metrics assessed. The dominant parameters are the channel erosion efficiencies (K_1 and K_2) and the erosion thresholds (ω_{c1} and ω_{c2}).

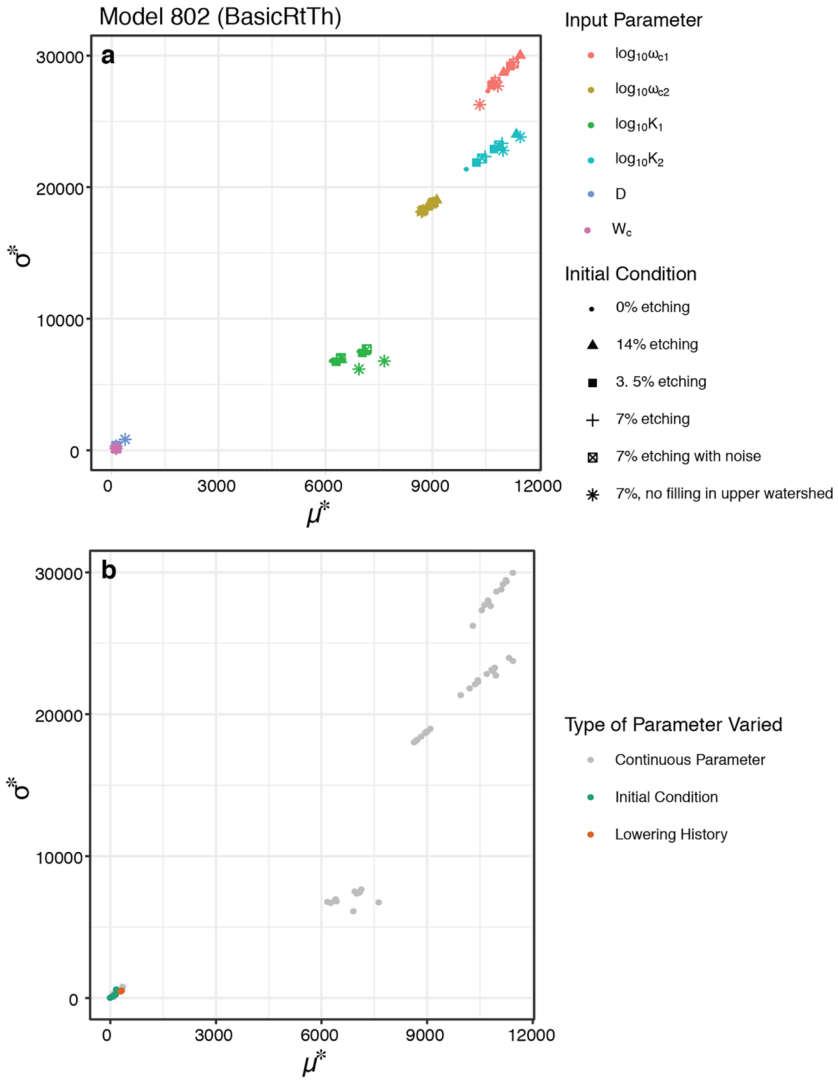

Figure 6. μ^* and σ^* for the objective function and model 802 BasicRtTh. This model was the best two-element model in the calibration effort presented by Barnhart et al. (2020). (a) The sensitivity of the objective function to changing the input parameter values and (b) the sensitivity to changing initial and boundary conditions. Figures presenting this information for all models are provided in the supporting information (Figures S6–S41).

This finding motivates the use of a single, averaged history in the calibration procedure. Similarly, the objective function also shows little sensitivity to reconstructed postglacial topography. The differences between the six different postglacial surfaces have little impact on the objective function. Based on this result, we use the initial condition with 7% etching and no change to bedrock in the upper portion of the watershed for the calibration in Barnhart et al. (2020b).

8.2.2. Model Parameters

Across the 36 models tested, sensitivity with respect to the objective function tends to be dominated by just a few parameters. MoM results for all models (Figures S6–S41) are similar to those shown in Figure 6. Parameters that control channel erosion consistently have high values for both μ^* and σ^*, which indicates that they have nonlinear importance for the simulated equivalents.

In Figure 7 we present the μ^* values for the eight models that differentiate between rock and till (the "Rt" series). This permutation is identified as the single most important based on calibration results (Barnhart et al., 2020b). The erodibility coefficients for river incision nearly always have a strong influence on model output, as measured by the objective function. For example, in the eight models illustrated in Figure 7, the erodibility coefficient parameters (K_1, K_2, K_{ss1}, and K_{ss2}) rank among the most influential for each model.

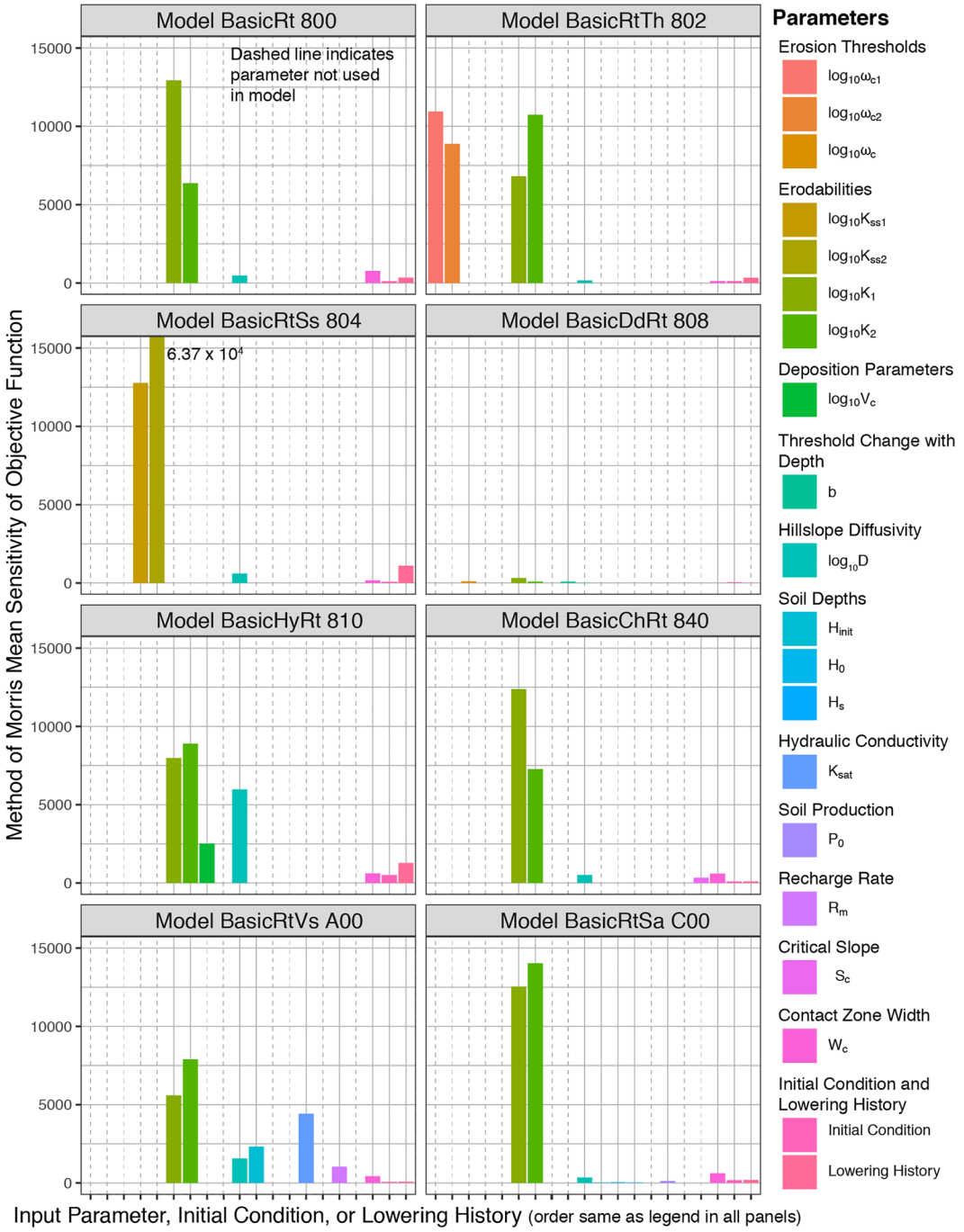

Figure 7. μ^* results for input parameters, initial conditions, and lowering histories for models with rock and till differentiation. Each panel shows results from a different model. For model parameters, the y axis shows the mean value of μ^* averaged across the categorical combination of six initial conditions and two lowering histories; for initial conditions and lowering histories, the y axis shows the mean value of μ^* averaged across the other categorical variable and parameter trajectory.

The hillslope transport efficiency coefficient, D, appears in all models. This parameter sets the rate at which hillslopes evolve. The parameter is the log-transformed coefficient, log D, and is rarely among the most influential parameters. This is the case both for the objective function (equation (4)), which purposefully placed emphasis on the channel regions of the model domain, and for the topographic metrics (Figure 5).

The critical slope parameter S_c (equation (S6)), which represents the gradient near and above which soil transport accelerates more-than-linearly with gradient, appears in models that use a nonlinear hillslope transport law (equation (S6)). In general, these models show relatively little sensitivity to S_c (see Figures S17, S31, and S38).

8.2.3. Spatial Variability in Sensitivity

The use of discrete landscape patches as components in the objective function (equation (4)) allows us to examine how each parameter influences the dynamics of the model with respect to a given patch.

As an illustration of how parameter sensitivity varies with location, Figure 8 shows sensitivity scores for each patch for model 802 BasicRtTh. The panels in Figure 8 are colored according to whether the patch lies primarily in the rock area (purple) or till area (orange) (see inset map). As expected, the areas within the till zone show strong sensitivity to the erodibility factor for till (K_1) and the erosion threshold for till (ω_{c1}), and little or no sensitivity to the corresponding parameters for rock. Conversely, in the upper portion of the watershed where bedrock dominates, the model shows strong sensitivity to rock erodibility K_2 and rock erosion threshold ω_{c2}.

The elements of the objective function that represent the patch scores (equation (2)) for the rock-dominated areas are sensitive to the parameters for till, especially to the till erosion threshold. This sensitivity reflects the fact that the rock-dominated area lies in the upper part of the watershed, where erosion rates and patterns depend in part on what is happening downstream. For example, when till is easy to erode in a particular model run (because K_1 is high or/and ω_{c1} is low), rapid erosion on the lower branches of the channel network will propagate upstream to the bedrock portion of the watershed, and thus tend to produce more rapid erosion there as well. The inferences drawn from the spatially distributed patches are broadly consistent with that from the drainage area and elevation measures of Figure 5.

9. Discussion

We can use the MoM sensitivity analysis to group the input parameters into three categories based on the μ^* and σ^* values:

1. low importance (small μ^*, small σ^*),
2. large, linear importance (large μ^*, small σ^*),
3. large, nonlinear importance (large μ^*, large σ^*).

Here we discuss the broader implications of the MoM results for landscape evolution model dynamics.

9.1. Sensitivity to Different Outputs

The results of the sensitivity analysis indicate that the output parameter of interest governs which input parameters are identified as most important. For example, the hillslope diffusivity D is not often an important parameter, but becomes important for the high percentile drainage area measures (Figure 5). Conversely, parameters that are often important are not important for every output parameter (e.g., the erosion thresholds ω_1 and ω_2 are not important to the low drainage area measures in Figure 5). Fluvial erodibility coefficients and thresholds are commonly important parameters. This is not a surprising finding, given that the erodibility coefficient governs the rate of water erosion under a given hydrologic scenario.

We find little sensitivity to initial or boundary conditions across all models. Prior work has demonstrated sensitivity to initial condition topography when a model is initialized with a low-relief surface decorated with random noise (Hancock, 2006; Hancock et al., 2016; Ijjász-Vásquez et al., 1992; Kwang & Parker, 2019; Perron & Fagherazzi, 2012; Willgoose et al., 2003). But in our case study, because the postglacial topography constrains a reasonably well-documented initial drainage pattern, such sensitivity to does not arise.

Evaluation of spatially variable sensitivity in two catchments by Skinner et al. (2018) found that sensitivity in most output metrics varied systematically with subcatchment stream order. We did not perform sensitivity analysis on multiple watersheds, but the topographic metrics and discretized cell-by-cell comparison both indicate that spatially variable outputs (e.g., Figure 8 and their Figure 5) show different sensitives.

Taken together, these results, along with those of Ziliani et al. (2013) and Skinner et al. (2018), do not point to universal findings in sensitivity analysis of landscape evolution models. It is reasonable to expect that

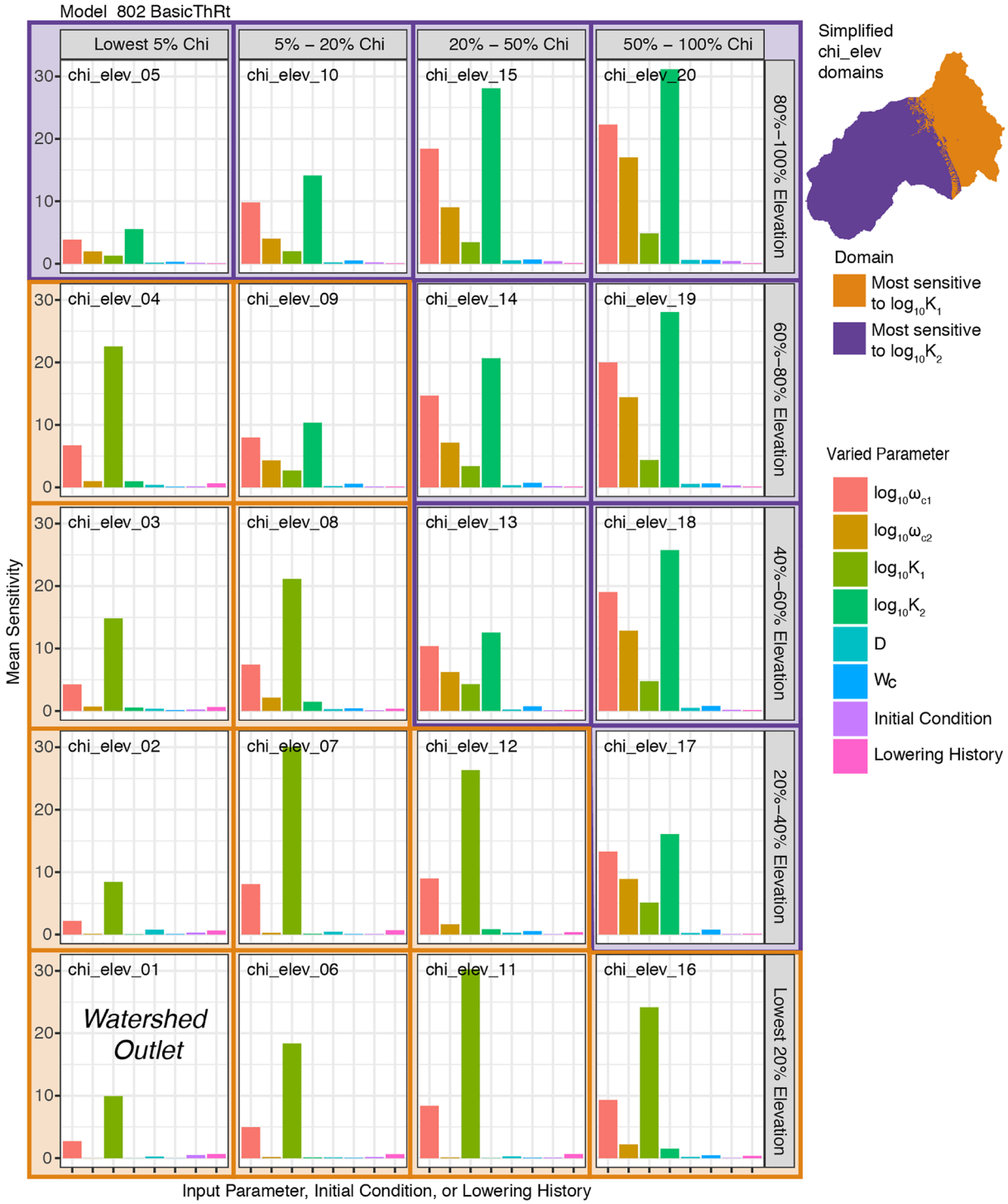

Figure 8. μ^* results for each landscape patch shown in Figure 4a different parameters, initial conditions, and lowering histories for models with rock and till differentiation. Each panel shows results from a different model. For model parameters, the y axis shows the mean value of μ^* averaged across the categorical combination of six initial conditions and two lowering histories; for initial conditions and lowering histories, the y axis shows the mean value of μ^* averaged across the other categorical variable and parameter trajectory.

sensitivity analysis results will reveal different parameter importance rankings for different models, different watersheds, and different outputs. Yet the results of sensitivity analysis need not be universal in order to be useful. Given the potential to add additional complexity, sensitivity analysis provides an approach to determine when that complexity changes outcomes relevant to a particular application.

9.2. Few Large, Linear Sensitivities

Large, linear sensitivities plot on the μ^*-σ^* plot in the lower right corner. These parameters are important, but do not interact with other parameters, such that the sensitivity is constant across parameter space. We find few parameters that demonstrate these properties. For example, the till erodibility K_1 plots toward the lower right corner in some models. These finding is consistent with prior MoM analyses in Earth surface process modeling where μ^*-σ^* are shown (Shobe et al., 2018; Skinner et al., 2018; Temme & Vanwalleghem, 2016; Ziliani et al., 2013).

Landscape evolution theory makes several predictions for linear sensitivities between model inputs and outputs. For example, a linear relationship between uplift rate and the relief of a mountain range is expected at topographic steady state when using a model with only fluvial incision based on stream power (Whipple & Tucker, 1999). Unlike this example, the case studies described here do not necessarily reach a steady state. This points to an additional dimension across which parameter sensitivities are likely to be inconsistent across applications: the duration of a model run in relation to characteristic adjustment timescales.

9.3. Parameter Importance Consistent With Model Physics

Finding that the patterns of parameter importance were inconsistent with model physics would be problematic. We find that sensitivities are in agreement with model physics. This is most intuitive in the context of the topographic metrics (Figure 5). We highlight this by focusing on the erodibility K_1, the thresholds ω_{c1} and ω_{c2}, and the diffusivity D.

The two thresholds ω_{c1} and ω_{c2} are the most important parameters for the 100th percentile of elevation (the maximum height in the catchment). This derives from these parameters' ability to effectively turn off erosion by channelized flow across portions of the landscape. In contrast the till erodibility K_1 is most important for the lowest percentile of elevation (2nd) consistent with its role setting the slope-area relationship of the lowest elevation portions of the channel network.

The diffusivity is important to the high percentiles of drainage area. These values are influenced by the details of how the main stem channels join. As the diffusivity impacts the movement of sediment from steeper hillslopes into the channels, its efficiency impacts channel network rearrangement, and thus drainage area percentiles.

9.4. Cross-Model Comparison

The most striking cross-model sensitivity difference comes from comparing the Th models that have constant erosion thresholds with the Dd models, which have depth-dependent erosion thresholds (e.g., BasicRtTh and BasicDdRt in Figure 7). The Dd models show very low sensitivity values compared to their Th counterparts.

These models, which were inspired by the hypothesis of Gran et al. (2013) that the effective erosion threshold for glacial till may increase progressively with incision depth, add a parameter, b, that describes the rate of erosion-threshold increase as incision progresses. By the definition of a model presented in section 2 the Dd models are distinct from the Th models in the addition of a degree of freedom—a value of $b = 0$ in a Dd model recreates the equivalent Th model. The dynamics of BasicRtTh in Figure 7 are present in the slice of BasicDdRt parameter space where $b = 0$. The low sensitivity of the BasicDdRt models reflects relatively high values of b: after a limited degree of incision, the threshold can become so high that erosion virtually halts.

This example highlights the importance of building complexity slowly (as recommended by Hill & Tiedeman, 2007). By considering BasicRtTh and BasicDdRt as separate models, and doing separate sensitivity analyses on them, we are able to identify the degree to which b influences this model. The observation that intermodel sensitivity—that is, the difference in sensitivity (or behavior more generally) between two seemingly similar models—is often large, is also in agreement with prior work. For example, Skinner et al. (2018) found that the choice of sediment transport formulae in CAESAR was often the most important model change.

9.5. The Challenge of Determining Parameter Ranges

Many of the parameters required by landscape evolution models have ranges spanning orders of magnitude, even when these are based on field observation, literature, and other observational data (see Text S5 and Barnhart et al., 2020c). In our application, a literature-based a priori parameter range for K, for example, led to model behavior that ranged from minimal erosion to unrealistically rapid removal of most of the topography in the study catchment. This raises the question, how to best determine reasonable parameter ranges?

One possible approach starts by asking: what are the reasonable limits for a "bad" objective function value? In our application, this would span from objective function values indicating virtually no erosion, to values associated with two or three times as much erosion as was expected to occur. After an initial estimation of parameter values is constructed, a mapping phase would identify the portion of parameter space in which the limits of a reasonable "bad" objection function value are not exceeded. Subsequent model analysis methods would then be restricted to this region of parameter space.

9.6. Relationship Between Sensitivity Analysis and Calibration

In the sensitivity analysis presented here we find that D is a relatively unimportant parameter. However, we find in our companion paper that the model structure permutation that includes nonlinear hillslope sediment transport improves model performance (Barnhart et al., 2020b). That D is not often important is not surprising, given that expected hillslope adjustment timescales (which for a linear model scale with the square of hillslope length, divided by D) are long relative to the timescale of our sensitivity analysis. It thus seems a contradiction that a process that depends on a relatively unimportant parameter is identified as important.

Here it is useful to distinguish between an important *process* and an important *parameter*. A trivial example is that of gravity. While the value for the gravitational constant g varies little on Earth, and studies of particle settling are relatively insensitive to the smaller decimals of g, the inclusion of g in equations for settling must include the process of gravitational attraction. Settling requires the existence of gravity, but its rate is insensitive to the (minimal) variation in g on Earth. Similarly, the results of the calibration indicate that the addition of the nonlinear hillslope transport *process* (replacing the default linear transport process) is important for model performance, but that the results are not particularly sensitive to the value of D. Further, they identify that it is only when this process interacts with a fluvial erosion threshold that its benefit is observed.

A reasonable argument could be made to not include hillslope diffusion in this effort based on an analysis of hillslope adjustment timescales and the duration of simulations. Yet this would mask identification of highly nonlinear interactions between model structure elements. This observation is not an argument that it is necessary to include all known physical processes. It is, however, an argument to consider plausible that processes operative on mismatched timescales (e.g., off by a factor of 10 or 100) may have unexpected interactions.

10. Conclusions

We develop a hierarchical set of landscape evolution models for application in a well-constrained site that has undergone rapid postglacial erosion. The model set efficiently explores the model structure space defined by twelve binary choices regarding the representation of geologic materials, hillslope sediment transport, surface-water hydrology, and erosion by channelized flow.

We outline an approach to mechanistic hypothesis testing in landscape evolution theory. The approach involves reconstructing paleotopography for a well-constrained site at a known point in geologic time, defining a set of alternative models, and applying methods of model analysis to the candidate model set. Here we presented results of the sensitivity analysis; a companion paper Barnhart et al. (2020b) describes model calibration.

Sensitivity analysis of topographic metrics as well as direct comparison of topography indicates that simulated equivalents are most sensitive to channel erosion parameters. In contrast, hillslope diffusivity is rarely influential. Unpacking our objective function to examine spatial patterns in sensitivity demonstrates that

model outputs designed to measure simulation performance in the upper portions of a watershed are sensitive to inputs that control the evolution of the lower portion of the watershed. We find that sensitivity to initial and boundary condition options is much smaller than sensitivity to parameter values. This reflects the high level of geologic constraint on both initial and boundary conditions at our study site.

Multimodel analysis techniques such as sensitivity analysis constitute both a tool and a philosophical approach. They provide a path forward to identify which parameters control landscape evolution model dynamics and which observables best constrain model structure.

Code and Data Availability

The creation and analysis of models presented in this three-part series was fully scripted. Instructions for reproducing the results (which took nearly 1 million core hours to run), input files, model and analysis code, and the model output files are available through a GlobusConnect endpoint (endpoint name: Barnhart_WVDP_EWG_STUDY3, endpoint identifier UUID 89df0600-bd11-11e8-8c12-0a1d4c5c824a). In addition, the input files and code are housed on GitHub (https://github.com/kbarnhart/inverting_topography_postglacial) and archived with Zenodo (Barnhart et al., 2020a).

Acknowledgments
Support for this work was provided by a contract with Enviro Compliance Solutions, Inc. (Contract DE-EM0002446/0920/13/DE-DT0005364/001), NSF Award 1450409 to Tucker, an NSF EAR Postdoctoral Fellowship to Barnhart (NSF 1725774), and a National Defense Science and Engineering Graduate Fellowship and a University of Colorado Chancellor's Fellowship to Shobe. Landlab is supported by NSF ACI-1450409 and by the Community Surface Dynamics Modeling System (CSDMS; NSF 1226297 and 1831623). This work utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (Awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University. We acknowledge computing time on the CU-CSDMS High-Performance Computing Cluster. Data storage is supported by the University of Colorado Boulder "PetaLibrary." We gratefully acknowledge Editor Amy East, Associate Editor Jon Pelletier, Tom Coulthard, and two anonymous reviewers for each considering all three manuscripts. Their constructive and thoughtful comments have substantially improved these manuscripts.

References

Adams, B. M., Bauman, L. E., Bohnhoff, W. J., Dalbey, K. R., Ebeida, M. S., Eddy, J. P., et al. (2017a). Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 6.6 theory manual (*Sandia Technical Report SAND2014-4253*).

Adams, B. M., Bauman, L. E., Bohnhoff, W. J., Dalbey, K. R., Ebeida, M. S., Eddy, J. P., et al. (2017b). Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 6.6 user manual (*Sandia Technical Report SAND2014-4253*).

Ahnert, F. (1976). Brief description of a comprehensive three-dimensional process-response model of landform development. *Zeitschrift fur Geomorfologie, Supplementband*, *25*, 29–49.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Anderson, R. S. (1994). Evolution of the Santa Cruz Mountains, California, through tectonic growth and geomorphic decay. *Journal of Geophysical Research*, *99*, 20,161–20,179. https://doi.org/10.1029/94JB00713

Andrews, D. J., & Bucknam, R. C. (1987). Fitting degradation of shoreline scarps by a nonlinear diffusion model. *Journal of Geophysical Research*, *92*(B12), 12,857–12,812. https://doi.org/10.1029/JB092iB12p12857

Andrews, D. J., & Hanks, T. C. (1985). Scarp degraded by linear diffusion: Inverse solution for age. *Journal of Geophysical Research*, *90*(B12), 10,193–10,208. https://doi.org/10.1029/JB090iB12p10193

Attal, M., Cowie, P., Whittaker, A., Hobley, D., Tucker, G. E., & Roberts, G. (2011). Testing fluvial erosion models using the transient response of bedrock rivers to tectonic forcing in the Apennines, Italy. *Journal of Geophysical Research*, *116*, F02005. https://doi.org/10.1029/2010JF001875

Barnhart, K. R., Glade, R. C., Shobe, C. M., & Tucker, G. E. (2019). Terrainbento 1.0: A Python package for model analysis in long-term drainage basin evolution. *Geoscientific Model Development*, *12*(4), 1267–1297. https://doi.org/10.5194/gmd-12-1267-2019

Barnhart, K. R., Hutton, E. W. H., & Tucker, G. E. (2019). umami: A Python package for Earth surface dynamics objective function construction. *Journal of Open Source Software*, *4*, 1776. https://doi.org/10.21105/joss.01776

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

Barnhart, K. R., Tucker, G. E., Doty, S., Shobe, C. M., Glade, R. C., Rossi, M. W., & Hill, M. C. (2020a). *Calculation package: Inverting topography for landscape evolution model process representation*. Zenodo. https://doi.org/10.5281/zenodo.2799489

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

Barnhart, K. R., Tucker, G. E., Doty, S., Shobe, C. M., Glade, R. C., Rossi, M. W., & Hill, M. C. (2020c). Inverting topography for landscape evolution model process representation: 3. Determining parameter ranges for select mature geomorphic transport laws and connecting changes in fluvial erodibility to changes in climate. *Journal of Geophysical Research: Earth Surface*, *125*, e2019JF005287. https://doi.org/10.1029/2019JF005287

Bennett, S. J. (2017). *Report of the West Valley Erosion Working Group study 2: Recent erosion and deposition processes: Task 2.2: Infiltration and soil moisture determination; Task 2.5: Erodibility of cohesive sediment; Task 2.6: Erodibility of clastic sediment in selected gullies, stream channels, and streambanks* (Tech. Rep.). Phase 1 Studies Erosion Working Group. Retrieved from https://wvphaseonestudies.emcbc.doe.gov/documents/EWG%20Study%202%20Report_Tasks%202.2,%202.5,%202.6_5.2.17.pdf

Beven, K., & Kirkby, M. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Journal*, *24*(1), 43–69.

Booth, A. M., & Roering, J. J. (2011). A 1-D mechanistic model for the evolution of earthflow-prone hillslopes. *Journal of Geophysical Research*, *116*, F04021. https://doi.org/10.1029/2011JF002024

Booth, A. M., Roering, F. J., & Rempel, A. W. (2013). Topographic signatures and a general transport law for deep-seated landslides in a landscape evolution model. *Journal of Geophysical Research: Earth Surface*, *118*, 603–624. https://doi.org/10.1002/jgrf.20051

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Verlag New York: Springer Science & Business Media.

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling and Software*, *22*(10), 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004

Chatanantavet, P., & Parker, G. (2009). Physically based modeling of bedrock incision by abrasion, plucking, and macroabrasion. *Journal of Geophysical Research*, *114*, F04018. https://doi.org/10.1029/2008JF001044

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*, W00B02. https://doi.org/10.1029/2007WR006735

Cortes, L. (2016). *2015 high resolution lidar & aerial orthoimagery, project western new york nuclear service, center and environs* (Tech. Rep.). Colorado Springs, CO, USA: The Sanborn Map Company, Inc.

Coulthard, T. J., Kirkby, M. J., & Macklin, M. G. (1998). Non-linearity and spatial resolution in a cellular automaton model of a small upland basin. *Hydrology and Earth System Sciences*, *2*(2/3), 257–264. https://doi.org/10.5194/hess-2-257-1998

Coulthard, T. J., Macklin, M., & Kirkby, M. (2002). A cellular model of holocene upland river basin and alluvial fan evolution. *Earth Surface Processes and Landforms*, *27*(3), 269–288. https://doi.org/10.1002/esp.318

Coulthard, T. J., Neal, J. C., Bates, P. D., Ramirez, J., de Almeida, G. A. M., & Hancock, G. R. (2013). Integrating the LISFLOOD-FP 2D hydrodynamic model with the CAESAR model: Implications for modelling landscape evolution. *Earth Surface Processes and Landforms*, *38*(15), 1897–1906. https://doi.org/10.1002/esp.3478

Coulthard, T. J., & Skinner, C. J. (2016). The sensitivity of landscape evolution models to spatial and temporal rainfall resolution. *Earth Surface Dynamics*, *4*(3), 757–771. https://doi.org/10.5194/esurf-4-757-2016

Croissant, T., & Braun, J. (2014). Constraining the stream power law: A novel approach combining a landscape evolution model and an inversion method. *Earth Surface Dynamics*, *2*(1), 155–166. https://doi.org/10.5194/esurf-2-155-2014

Culling, W. (1963). Soil creep and the development of hillside slopes. *The Journal of Geology*, *71*, 127–161. Retrieved from http://www.jstor.org/stable/30066150

Davy, P., & Lague, D. (2009). Fluvial erosion/transport equation of landscape evolution models revisited. *Journal of Geophysical Research*, *114*, F03007. https://doi.org/10.1029/2008JF001146

Deal, E., Braun, J., & Botter, G. (2018). Understanding the role of rainfall and hydrology in determining fluvial erosion efficiency. *Journal of Geophysical Research: Earth Surface*, *123*, 744–778. https://doi.org/10.1002/2017JF004393

Densmore, A. L., Ellis, M. A., & Anderson, R. S. (1998). Landsliding and the evolution of normal-fault-bounded mountains. *Journal of Geophysical Research*, *103*(B7), 15,203–15,219. https://doi.org/10.1029/98JB00510

DiBiase, R. A., & Whipple, K. X. (2011). The influence of erosion thresholds and runoff variability on the relationships among topography, climate, and erosion rate. *Journal of Geophysical Research*, *116*, F04036. https://doi.org/10.1029/2011JF002095

Dietrich, WE, Bellugi, D. G., Sklar, L. S., Stock, J. D., Heimsath, AM, & Roering, J. J. (2003). Geomorphic transport laws for predicting landscape form and dynamics. In P. Wilcock & R. Iverson (Eds.), *Prediction in geomorphology*, Geophysical Monograph Series. Washington, DC: AGU. https://doi.org/10.1029/135GM09

Dietrich, W. E., Wilson, C. J., Montgomery, D. R., & McKean, J. (1993). Analysis of erosion thresholds, channel networks, and landscape morphology using a digital terrain model. *The Journal of Geology*, *101*(2), 259–278. https://doi.org/10.1086/648220

Doane, T. H., Furbish, D. J., Roering, J. J., Schumer, R., & Morgan, D. J. (2018). Nonlocal sediment transport on steep lateral moraines, Eastern Sierra Nevada, California, USA. *Journal of Geophysical Research: Earth Surface*, *123*, 187–208. https://doi.org/10.1002/2017JF004325

Dunne, T. (1970). Runoff production in a humid area, U.S. Agricultural Research Service.

Dunne, T., & Black, R. D. (1970). Partial area contributions to storm runoff in a small new england watershed. *Water Resources Research*, *6*(5), 1296–1311. https://doi.org/10.1029/WR006i005p01296

ESRI (2014). *Arcgis desktop: Release 10.3*.

Einstein, H. A. (1950). The bed-load function for sediment transportation in open channel flows (Tech. Rep. No. 1026). Washington, D.C., USA: United States Department of Agriculture, Soil Conservation Service.

Fakundiny, R. (1985). Practical applications of geological methods at the west valley low-level radioactive waste burial ground, western New York. *Northeastern Environmental Science*, *4*(3-4), 116–148.

Foglia, L., Mehl, S., Hill, M., & Burlando, P. (2013). Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland. *Water Resources Research*, *49*, 260–282. https://doi.org/10.1029/2011WR011779

Fox, M., Goren, L., May, D., & Willett, S. (2014). Inversion of fluvial channels for paleorock uplift rates in Taiwan. *Journal of Geophysical Research: Earth Surface*, *119*, 1853–1875. https://doi.org/10.1002/2014JF003196

Furbish, D. J. (2003). Using the dynamically coupled behavior of land-surface geometry and soil thickness in developing and testing hillslope evolution models, *Prediction in geomorphology* (pp. 169–181). Washington, DC: American Geophysical Union (AGU). https://doi.org/10.1029/135GM12

Gilbert, G. K. (1909). The convexity of hilltops. *The Journal of Geology*, *17*(4), 344–350. https://doi.org/10.1086/621620

Goren, L., Fox, M., & Willett, S. (2014). Tectonics from fluvial topography using formal linear inversion: Theory and applications to the Inyo Mountains, California. *Journal of Geophysical Research: Earth Surface*, *119*, 1651–1681. https://doi.org/10.1002/2014JF003079

Gran, K. B., Finnegan, N., Johnson, A. L., Belmont, P., Wittkop, C., & Rittenour, T. (2013). Landscape evolution, valley excavation, and terrace development following abrupt postglacial base-level fall. *Geological Society of America Bulletin*, *125*(11-12), 1851–1864. https://doi.org/10.1130/B30772.1

Gray, H. J., Shobe, C. M., Hobley, D. E. J., Tucker, G. E., Duvall, A. R., Harbert, S. A., & Owen, L. A. (2018). Off-fault deformation rate along the southern San Andreas fault at Mecca Hills, Southern California, inferred from landscape modeling of curved drainages. *Geology*, *46*(1), 59–62. https://doi.org/10.1130/G39820.1

Hancock, G. R. (2006). The impact of different gridding methods on catchment geomorphology and soil erosion over long timescales using a landscape evolution model. *Earth Surface Processes and Landforms*, *31*(8), 1035–1050. https://doi.org/10.1002/esp.1306

Hancock, G. R., & Coulthard, T. J. (2012). Channel movement and erosion response to rainfall variability in southeast Australia. *Hydrological Processes*, *26*(5), 663–673. https://doi.org/10.1002/hyp.8166

Hancock, G. R., Coulthard, T. J., & Lowry, J. B. C. (2016). Predicting uncertainty in sediment transport and landscape evolution—The influence of initial surface conditions. *Computers and Geosciences*, *90*(Part B), 117–130. https://doi.org/10.1016/j.cageo.2015.08.014

Hancock, G. R., Coulthard, T. J., Martinez, C., & Kalma, J. D. (2011). An evaluation of landscape evolution models to simulate decadal and centennial scale soil erosion in grassland catchments. *Journal of Hydrology*, *398*(3-4), 171–183. https://doi.org/10.1016/j.jhydrol.2010.12.002

Hancock, G. R., Lowry, J. B. C., & Coulthard, T. J. (2015). Catchment reconstruction—Erosional stability at millennial time scales using landscape evolution models. *Geomorphology*, *231*(C), 15–27. https://doi.org/10.1016/j.geomorph.2014.10.034

Hancock, G. R., Lowry, J., Coulthard, T. J., Evans, K., & Moliere, D. (2010). A catchment scale evaluation of the SIBERIA and CAESAR landscape evolution models. *Earth Surface Processes and Landforms*, *35*(8), 863–875. https://doi.org/10.1002/esp.1863

Hancock, G. R., Wells, T., Dever, C., & Braggins, M. (2018). Hillslope and point based soil erosion—An evaluation of a landscape evolution model. *Earth Surface Processes and Landforms*, *44*(5), 1163–1177. https://doi.org/10.1002/esp.4566

Hancock, G. R., & Willgoose, G. R. (2001). Use of a landscape simulator in the validation of the SIBERIA catchment evolution model: Declining equilibrium landforms. *Water Resources Research*, *37*(7), 1981–1992. https://doi.org/10.1029/2001WR900002

Hancock, G. R., Willgoose, G. R., & Evans, K. G. (2002). Testing of the SIBERIA landscape evolution model using the Tin Camp Creek, Northern Territory, Australia, field catchment. *Earth Surface Processes and Landforms*, *27*(2), 125–143. https://doi.org/10.1002/esp.304

Hanks, T. C. (2000). The age of scarplike landforms from diffusion-equation analysis. In *Quaternary Geochronology: Methods and Applications* (pp. 313–338). https://doi.org/10.1029/RF004p0313

Harkins, N., Kirby, E., Heimsath, A. M., Robinson, R., & Reiser, U. (2007). Transient fluvial incision in the headwaters of the Yellow River, northeastern Tibet. *Journal of Geophysical Research*, *112*, F03S04. https://doi.org/10.1029/2006JF000570

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (1997). The soil production function and landscape equilibrium. *Nature*, *388*(6640), 358–361. https://doi.org/10.1038/41056

Herman, F., & Braun, J. (2006). A parametric study of soil transport mechanisms, *Special paper 398: Tectonics, climate, and landscape evolution* (pp. 191–200). Boulder, CO, USA: Geological Society of America. https://doi.org/10.1130/2006.2398(11)

Hill, M. C., & Tiedeman, C. R. (2007). *Effective groundwater model calibration: With analysis of data, sensitivities, predictions, and uncertainty*. Hoboken, NJ, USA: John Wiley. https://doi.org/10.1002/0470041080

Hobley, D. E., Adams, J. M., Nudurupati, S. S., Hutton, E. W., Gasparini, N. M., Istanbulluoglu, E., & Tucker, G. E. (2017). Creative computing with landlab: An open-source toolkit for building, coupling, and exploring two-dimensional numerical models of Earth-surface dynamics. *Earth Surface Dynamics*, *5*(1), 21–46. https://doi.org/10.5194/esurf-5-21-2017

Hobley, D. E., Sinclair, H. D., Mudd, S. M., & Cowie, P. A. (2011). Field calibration of sediment flux dependent river incision. *Journal of Geophysical Research*, *116*, F04017. https://doi.org/10.1029/2010JF001935

Howard, A. D. (1994). A detachment-limited model of drainage basin evolution. *Water Resources Research*, *30*(7), 2261–2285. https://doi.org/10.1029/94WR00757

Howard, A. D., & Kerby, G. (1983). Channel changes in badlands. *Geological Society of America Bulletin*, *94*, 739–752. https://doi.org/10.1130/0016-7606(1983)94<739:CCIB>2.0.CO;2

Howard, A. D., & Tierney, H. E. (2012). Taking the measure of a landscape: Comparing a simulated and natural landscape in the Virginia coastal plain. *Geomorphology*, *137*(1), 27–40. https://doi.org/10.1016/j.geomorph.2010.09.031

Ibbitt, R. P., Willgoose, G. R., & Duncan, M. J. (1999). Channel network simulation models compared with data from the Ashley River, New Zealand. *Water Resources Research*, *35*(12), 3875–3890. https://doi.org/10.1029/1999WR900245

Ijjász-Vásquez, E. J., Bras, R. L., & Moglen, G. E. (1992). Sensitivity of a basin evolution model to the nature of runoff production and to initial conditions. *Water Resources Research*, *28*(10), 2733–2741. https://doi.org/10.1029/92WR01561

Jansen, M. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, *117*(1-2), 35–43. https://doi.org/10.1016/S0010-4655(98)00154-4

Kirkby, M., & Bull, L. (2000). Some factors controlling gully growth in fine-grained sediments: A model applied in southeast Spain. *Catena*, *40*(2), 127–146. https://doi.org/10.1016/S0341-8162(99)00077-6

Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*, 89–97. https://doi.org/10.5194/adgeo-5-89-2005

Kwang, J. S., & Parker, G. (2019). Extreme memory of initial conditions in numerical landscape evolution models. *Geophysical Research Letters*, *46*, 6563–6573. https://doi.org/10.1029/2019GL083305

LaFleur, R. G. (1979). *Glacial geology and stratigraphy of western New York nuclear service center and vicinity, Cattaraugus and Erie counties, New York* (Tech. Rep. No. Open-File Report 79-989). Albany, NY, USA: US Geological Survey. https://doi.org/10.3133/ofr79989

Lague, D., Hovius, N., & Davy, P. (2005). Discharge, discharge variability, and the bedrock channel profile. *Journal of Geophysical Research*, *110*, F04006. https://doi.org/10.1029/2004JF000259

Liu, Z., Otto-Bliesner, B. L., He, F., Brady, E. C., Tomas, R., Clark, P. U., et al. (2009). Transient simulation of last deglaciation with a new mechanism for Bølling-Allerød warming. *Science*, *325*(5938), 310–314. https://doi.org/10.1126/science.1171041

Loget, N., Davy, P., & Van Den Driessche, J. (2006). Mesoscale fluvial erosion parameters deduced from modeling the Mediterranean sea level drop during the Messinian (Late Miocene). *Journal of Geophysical Research*, *111*, F03005. https://doi.org/10.1029/2005JF000387

McGuire, L. A., Rengers, F. K., Kean, J. W., & Staley, D. M. (2017). Debris flow initiation by runoff in a recently burned basin: Is grain-by-grain sediment bulking or en masse failure to blame? *Geophysical Research Letters*, *44*, 7310–7319. https://doi.org/10.1002/2017GL074243

Milliman, J. D., & Syvitski, P. M. (1992). Geomorphic/tectonic control of sediment transport to the ocean: The importance of small mountainous rivers. *Journal of Geology*, *100*, 525–544. https://doi.org/10.1086/629606

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174. https://doi.org/10.2307/1269043

O'Loughlin, E. (1986). Prediction of surface saturation zones in natural catchments by topographic analysis. *Water Resources Research*, *22*(5), 794–804. https://doi.org/10.1029/WR022i005p00794

Passalacqua, P., Belmont, P., Staley, D. M., Simley, J. D., Arrowsmith, J. R., Bode, C. A., et al. (2015). Analyzing high resolution topography for advancing the understanding of mass and energy transfer through landscapes: A review. *Earth Science Reviews*, *148*(C), 174–193. https://doi.org/10.1016/j.earscirev.2015.05.012

Pelletier, J. D. (2007). Numerical modeling of the Cenozoic geomorphic evolution of the southern Sierra Nevada, California. *Earth and Planetary Science Letters*, *259*(1-2), 85–96. https://doi.org/10.1016/j.epsl.2007.04.030

Pelletier, J. D., DeLong, S. B., Al-Suwaidi, A. H., Cline, M., Lewis, Y., Psillas, J. L., & Yanites, B. (2006). Evolution of the Bonneville shoreline scarp in west-central Utah: Comparison of scarp-analysis methods and implications for the diffusion model of hillslope evolution. *Geomorphology*, *74*, 257–270. https://doi.org/10.1016/j.geomorph.2005.08.008

Pelletier, J. D., McGuire, L. A., Ash, J. L., Engelder, T. M., Hill, L. E., Leroy, K. W., et al. (2011). Calibration and testing of upland hillslope evolution models in a dated landscape: Banco Bonito, New Mexico. *Journal of Geophysical Research*, *116*, F04004. https://doi.org/10.1029/2011JF001976

Perera, H. J., & Willgoose, G. R. (1998). A physical explanation of the cumulative area distribution curve. *Water Resources Research*, *34*(5), 1335–1343. https://doi.org/10.1029/98WR00259

Perron, J. T., & Fagherazzi, S. (2012). The legacy of initial conditions in landscape evolution. *Earth Surface Processes and Landforms*, *37*(1), 52–63. https://doi.org/10.1002/esp.2205

Perron, J. T., & Royden, L. (2013). An integral approach to bedrock river profile analysis. *Earth Surface Processes and Landforms*, *38*(6), 570–576. https://doi.org/10.1002/esp.2205

Petit, C., Gunnell, Y., Saholiariliva, N. G., Meyer, B., & Séguinot, J. (2009). Faceted spurs at normal fault scarps: Insights from numerical modeling. *Journal of Geophysical Research*, *114*, B05403. https://doi.org/10.1029/2008JB005955

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling and Software*, *79*(C), 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008

Poeter, E. P., & Hill, M. C. (2007). *MMA, a computer code for multi-model analysis* (Tech. Rep. No. Techniques and Methods 6-E3). Boulder, CO, USA: United States Geological Survey. https://doi.org/10.3133/tm6E3

Portenga, E. W., & Bierman, P. R. (2011). Understanding Earth's eroding surface with ^{10}Be. *GSA Today*, *21*(8), 4–10. https://doi.org/10.1130/G111A.1

Pritchard, D., Roberts, G. G., White, N. J., & Richardson, C. N. (2009). Uplift histories from river profiles. *Geophysical Research Letters*, *36*, L24301. https://doi.org/10.1029/2009GL040928

Randall, A. D. (1980). *Glacial stratigraphy in part of Buttermilk Creek Valley: In Lafleur, R.G., Guidebook for 43rd annual reunion for Northeast Friends of Pleistocene* (Tech. Rep.). Troy, NY, USA: Northeast Friends of Pleistocene.

Rengers, F. K., & Tucker, G. E. (2014). Analysis and modeling of gully headcut dynamics, North American high plains. *Journal of Geophysical Research: Earth Surface*, *119*, 983–1003. https://doi.org/10.1002/2013JF002962

Roberts, G. G., & White, N. (2010). Estimating uplift rate histories from river profiles using African examples. *Journal of Geophysical Research*, *115*, B02406. https://doi.org/10.1029/2009JB006692

Roberts, G. G., White, N. J., Martin-Brandis, G. L., & Crosby, A. G. (2012). An uplift history of the Colorado Plateau and its surroundings from inverse modeling of longitudinal river profiles. *Tectonics*, *31*, TC4022. https://doi.org/10.1029/2012TC003107

Roe, G. H., Whipple, K. X., & Fletcher, J. K. (2008). Feedbacks among climate, erosion, and tectonics in a critical wedge orogen. *American Journal of Science*, *308*(7), 815–842. https://doi.org/10.2475/07.2008.01

Roering, J. J. (2008). How well can hillslope evolution models "explain" topography? Simulating soil transport and production with high-resolution topographic data. *Geological Society of America Bulletin*, *120*(9-10), 1248–1262. https://doi.org/10.1130/B26283.1

Rosenbloom, N. A., & Anderson, R. S. (1994). Hillslope and channel evolution in a marine terraced landscape, Santa Cruz, California. *Journal of Geophysical Research*, *99*, 14,013–14,029. https://doi.org/10.1029/94JB00048

Saltelli, A., & Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling and Software*, *25*(12), 1508–1517. https://doi.org/10.1016/j.envsoft.2010.04.012

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global sensitivity analysis: The primer*. Chichester, England; Hoboken, NJ: John Wiley. https://doi.org/10.1002/9780470725184

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models*. New York, NY: John Wiley.

Scheidt, C., Li, L., & Caers, J. (2018). *Quantifying uncertainty in subsurface systems* (Vol. 236). New York, NY: John Wiley. https://doi.org/10.1002/9781119325888

Shobe, C. M., Tucker, G. E., & Barnhart, K. R. (2017). The SPACE 1.0 model: A Landlab component for 2-D calculation of sediment transport, bedrock erosion, and landscape evolution. *Geoscientific Model Development*, *10*(12), 4577–4604. https://doi.org/10.5194/gmd-10-4577-2017

Shobe, C. M., Tucker, G. E., & Rossi, M. W. (2018). Variable-threshold behavior in rivers arising from hillslope-derived blocks. *Journal of Geophysical Research: Earth Surface*, *123*, 1931–1957. https://doi.org/10.1029/2017JF004575

Sidorchuk, A. (1999). Dynamic and static models of gully erosion. *Catena*, *37*(3), 401–414. https://doi.org/10.1016/S0341-8162(99)00029-6

Skinner, C. J., Coulthard, TJ, Schwanghart, W., Van De Wiel, M. J., & Hancock, G. R. (2018). Global sensitivity analysis of parameter uncertainty in landscape evolution models. *Geoscientific Model Development*, *11*(12), 4873–4888. https://doi.org/10.5194/gmd-11-4873-2018

Sklar, L. S., & Dietrich, W. E. (2004). A mechanistic model for river incision into bedrock by saltating bed load. *Water Resources Research*, *40*, W06301. https://doi.org/10.1029/2003WR002496

Small, E. E., Anderson, R. S., & Hancock, G. S. (1999). Estimates of the rate of regolith production using ^{10}Be and ^{26}Al from an alpine hillslope. *Geomorphology*, *27*(1), 131–150. https://doi.org/10.1016/S0169-555X(98)00094-4

Snyder, N. P., Whipple, K. X., Tucker, G. E., & Merritts, D. J. (2003). Importance of a stochastic distribution of floods and erosion thresholds in the bedrock river incision problem. *Journal of Geophysical Research*, *108*(B2), 2117. https://doi.org/10.1029/2001JB001655

Sweeney, K. E., & Roering, J. J. (2017). Rapid fluvial incision of a late Holocene lava flow: Insights from LiDAR, alluvial stratigraphy, and numerical modeling. *GSA Bulletin*, *129*(3-4), 500–512. https://doi.org/10.1130/B31537.1

Tarantola, A. (1987). *Inverse problem theory and methods for model parameter estimation*. Paris, France: Elsevier Science Pub. Co.

Tarantola, A., & Valette, B. J. (1982). Inverse problems quest for information. *Journal of Geophysics*, *50*, 159–170. Retrieved from https://journal.geophysicsjournal.com/JofG/article/view/28

Temme, A., & Vanwalleghem, T. (2016). LORICA—A new model for linking landscape and soil profile evolution: Development and sensitivity analysis. *Computers & Geosciences*, *90*, 131–143. Uncertainty and Sensitivity in Surface Dynamics Modeling https://doi.org/10.1016/j.cageo.2015.08.004

Tomkin, J. H., Brandon, M. T., Pazzaglia, F. J., Barbour, J. R., & Willett, S. D. (2003). Quantitative testing of bedrock incision models for the Clearwater River, NW Washington State. *Journal of Geophysical Research*, *108*(B6), 2308. https://doi.org/10.1029/2001JB000862

Tucker, G. E. (2004). Drainage basin sensitivity to tectonic and climatic forcing: Implications of a stochastic model for the role of entrainment and erosion thresholds. *Earth Surface Processes and Landforms*, *29*(2), 185–205. https://doi.org/10.1002/esp.1020

Tucker, G. E. (2009). Natural experiments in landscape evolution. *Earth Surface Processes and Landforms*, *34*, 1450–1460. https://doi.org/10.1002/esp.1833

Tucker, G. E., & Bras, R. L. (2000). A stochastic approach to modeling the role of rainfall variability in drainage basin evolution. *Water Resources Research*, *36*(7), 1953–1964. https://doi.org/10.1029/2000WR900065

Tucker, G. E., & Hancock, G. R. (2010). Modelling landscape evolution. *Earth Surface Processes and Landforms*, *46*, 28–50. https://doi.org/10.1002/esp.1952

Valla, P. G., van der Beek, P. A., & Lague, D. (2010). Fluvial incision into bedrock: Insights from morphometric analysis and numerical modeling of gorges incising glacial hanging valleys (Western Alps, France). *Journal of Geophysical Research*, *115*, F02010. https://doi.org/10.1029/2008JF001079

van der Beek, P., & Bishop, P. (2003). Cenozoic river profile development in the Upper Lachlan catchment (SE Australia) as a test quantitative fluvial incision models. *Journal of Geophysical Research*, *108*(b6), 2309. https://doi.org/10.1029/2002JB002125

Whipple, K. X. (2004). Bedrock rivers and the geomorphology of active orogens. *Annual Review of Earth and Planetary Sciences*, *32*, 151–185. https://doi.org/10.1146/annurev.earth.32.101802.120356

Whipple, K. X., & Tucker, G. E. (1999). Dynamics of the stream-power river incision model: Implications for height limits of mountain ranges, landscape response timescales, and research needs. *Journal of Geophysical Research*, *104*, 17,661–17,674. https://doi.org/10.1029/1999JB900120

Whipple, K. X., & Tucker, G. E. (2002). Implications of sediment-flux-dependent river incision models for landscape evolution. *Journal of Geophysical Research*, *107*, ETG 3–1-ETG 3-20. https://doi.org/10.1029/2000JB000044

Wilcock, P. R., & Crowe, J. C. (2003). Surface-based transport model for mixed-size sediment. *Journal of Hydraulic Engineering*, *129*(2), 120–128. https://doi.org/10.1061/(ASCE)0733-9429(2003)129:2(120)

Willenbring, J. K., Codilean, A. T., & McElroy, B. (2013). Earth is (mostly) flat: Apportionment of the flux of continental sediment over millennial time scales. *Geology*, *41*(3), 343–346. https://doi.org/10.1130/G33918.1

Willgoose, G. R., Bras, R. L., & Rodriguez-Iturbe, I. (1991a). A coupled channel network growth and hillslope evolution model: 1. Theory. *Water Resources Research*, *27*(7), 1671–1684. https://doi.org/10.1029/91WR00935

Willgoose, G. R., Bras, R. L., & Rodriguez-Iturbe, I. (1991b). A coupled channel network growth and hillslope evolution model: 2. Non-dimensionalization and applications. *Water Resources Research*, *27*(7), 1685–1696. https://doi.org/10.1029/91WR00936

Willgoose, G. R., Hancock, G. R., & Kuczera, G. (2003). A framework for the quantitative testing of landform evolution models, *Prediction in geomorphology* (pp. 195–216). Washington, DC: American Geophysical Union (AGU). https://doi.org/10.1029/135GM14

Wilson, M., & Young, R. A. (2018). *Phase 1 erosion studies: Study 1—Terrain analysis* (Tech. Rep.). West Valley Erosion Working Group. Retrieved from http://www.westvalleyphaseonestudies.org/Documents/EWG%20Final%20Study%201%20Report_Vol%201_2.21.18.pdf

Zhang, W., & Montgomery, D. R. (1994). Digital elevation model grid size, landscape representation, and hydrologic simulations. *Water Resources Research*, *30*(4), 1019–1028. https://doi.org/10.1029/93WR03553

Ziliani, L., Surian, N., Coulthard, T. J., & Tarantola, S. (2013). Reduced-complexity modeling of braided rivers: Assessing model performance by sensitivity analysis, calibration, and validation. *Journal of Geophysical Research: Earth Surface*, *118*, 2243–2262. https://doi.org/10.1002/jgrf.20154