JGR Earth Surface

RESEARCH ARTICLE
10.1029/2018JF004963

Inverting Topography for Landscape Evolution Model Process Representation: 2. Calibration and Validation

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

This article is a companion to Barnhart et al. (2020b, 2020c), https://doi.org/10.1029/2018JF004961 and https://doi.org/10.1029/2019JF005287.

Key Points:
- Comparison of calibrated models at a well-constrained field site serves as a formal hypothesis test of landscape evolution theory
- Representing subtle lithologic contrasts, fluvial incision thresholds, and nonlinear hillslope transport improves simulation results most
- Model structure improvements are not always linearly additive

Supporting Information:
- Supporting Information S1

Correspondence to:
K. R. Barnhart,
barnhark@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: 2. Calibration and validation. *Journal of Geophysical Research: Earth Surface*, *125*, e2018JF004963. https://doi.org/10.1029/2018JF004963

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

Abstract We present a multimodel analysis for mechanistic hypothesis testing in landscape evolution theory. The study site is a watershed with well-constrained initial and boundary conditions in which a river network locally incised 50 m over the last 13 ka. We calibrate and validate a set of 37 landscape evolution models designed to hierarchically test elements of complexity from four categories: hillslope processes, channel processes, surface hydrology, and representation of geologic materials. Comparison of each model to a base model, which uses stream power channel incision, uniform lithology, hillslope transport by linear diffusion, and surface water discharge proportional to drainage area, serves as a formal test of which elements of complexity improve model performance. Model fit is assessed using an objective function based on a direct difference between observed and simulated modern topography. A hybrid optimization scheme identifies optimal parameters and uncertainty. Multimodel analysis determines which elements of complexity improve simulation performance. Validation tests which model improvements persist when models are applied to an independent watershed. The three most important model elements are (1) spatial variation in lithology (differentiation between shale and glacial till), (2) a fluvial erosion threshold, and (3) a nonlinear relationship between slope and hillslope sediment flux. Due to nonlinear interactions between model elements, some process representations (e.g., nonlinear hillslopes) only become important when paired with the inclusion of other processes (e.g., erosion thresholds). This emphasizes the need for caution in identifying the minimally sufficient process set. Our approach provides a general framework for hypothesis testing in landscape evolution.

1. Introduction

Earth's surface exhibits a diversity of landforms, each shaped through the interaction of climate (including the roles of liquid water, ice, and wind), tectonics, and material properties of the surface (including the roles of lithology, rock fractures, and soil formation). An important open question in quantitative geomorphology is the extent to which these landforms contain information about their genesis (Davis, 1892; Gilbert, 1877, 1909). A more formal statement of this question is as follows: To what extent do fundamental observable quantities (such as topography and metrics derived from it) encode information about the equations that govern Earth surface evolution? This question is challenging to assess because we still lack (1) generalized governing equations that are readily assessed for any landscape from surface topography alone and (2) agreed-upon approaches for comparing observations and models (Dietrich et al., 2003; Hancock et al., 2010, 2011; Hancock & Willgoose, 2001; Howard & Tierney, 2012; Ibbitt et al., 1999; Perera & Willgoose, 1998; Skinner et al., 2018). In some landscapes, this challenge can be overcome by making reasonable simplifications to well-vetted governing equations that can be implemented on geologic time and space scales. Here we use formal model analysis to interrogate the properties of common variants of landscape evolution models and to identify what elements of additional complexity in model structure improve (or fail to improve) simulation performance. Specifically, we independently calibrate alternative landscape evolution models and assess when adding complexity or changing model structure improves simulation performance.

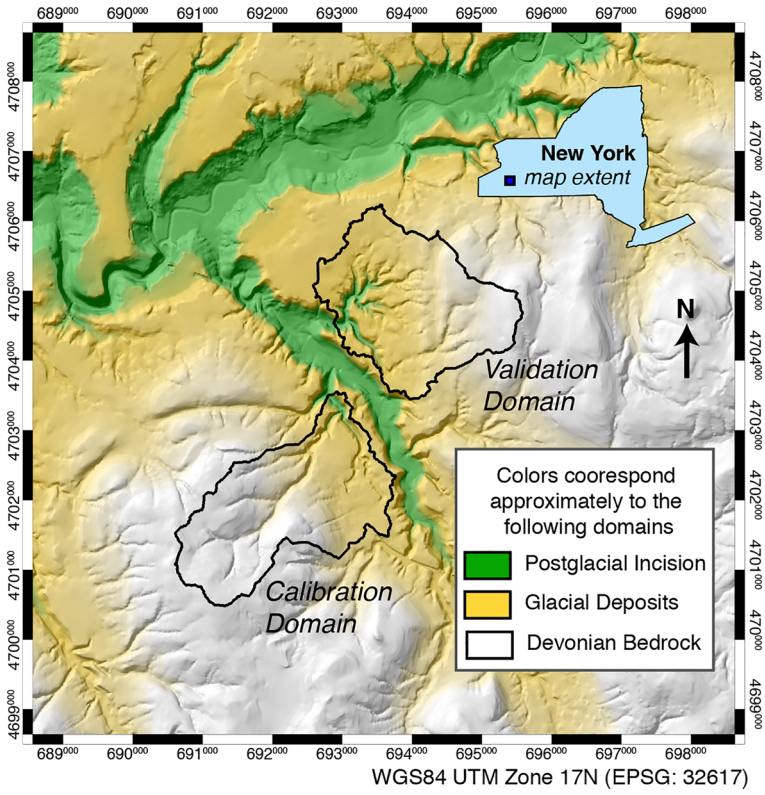

Figure 1. Hillshade of the study area indicating the extent of the calibration and validation domains used by Barnhart et al. (2020b). The three map colors indicate the three major geomorphic domains of the region. Topographic data from the National Elevation Database.

To test and calibrate alternative models, we take advantage of a well-constrained case study in postglacial landscape evolution, where remnants of a once continuous surface make it possible to reconstruct the latest Pleistocene paleotopography. We focus on the Franks Creek watershed, a ~5 km^2 drainage basin in western New York State, USA (Figure 1, catchment labeled "Calibration Domain"). The site is described by Barnhart et al. (2020b, their section 3); here we summarize some of its salient aspects. The watershed, which is underlain by a combination Devonian shale (Buehler & Tesmer, 1963) and late Pleistocene glacial sediments, was glaciated until roughly 13 ka (Wilson & Young, 2018). Both the topographic surface at deglaciation and the incision history of the channel downstream of the study watershed are well constrained (Fakundiny, 1985; Wilson & Young, 2018), which makes the watershed a good natural experiment with which to test landscape evolution models (Tucker, 2009). A second nearby watershed with similar size, relief, morphology, and base level incision history serves as an independent validation site for calibrated models (Figure 1, catchment labeled "Validation Domain").

We use the study watersheds shown in Figure 1 as the basis for analyzing, testing, and calibrating a series of alternative models of long-term landscape evolution, using an approach to model analysis that is outlined in a companion paper (Barnhart et al., 2020b, their section 2). The choice of models is summarized in section 2, and presented in greater detail by Barnhart et al. (2020b, their section 4). The *objective function* that is used as quantitative basis for assessing model performance is summarized in section 3.3, and also described in greater detail by Barnhart et al. (2020b, their section 6).

Because the term *model* is used in many different ways in the natural sciences (e.g., Bras et al., 2003), it is useful to be precise about the meaning of the word in the context of this study. Here we summarize the more extensive discussion in Barnhart et al. (2020b, their section 2). We use the term model generally to describe any quantitative link between conceptual or theoretical descriptions of a system and *simulated equivalents* of observable data. A model is therefore any analytical or numerical expression relating input parameters to simulated equivalents. To appreciate this definition, consider a simple model that fails to capture the internal dynamics of a study system. If one wanted to improve its performance, one might either adjust

the value of a given *parameter*, or make more fundamental changes in *model structure*. A change in model structure implies a change in the mathematical form of one or more governing equations, thereby making it a new and different model. A change in model structure may also change the set of state variables that are represented or a change in the number of degrees of freedom (e.g., by converting a free parameter to a fixed value). Other studies have used different definitions of a model (e.g., Pianosi et al., 2016; Skinner et al., 2018). Our working definition is complementary to our goal: to formally test when adding complexity adds to simulation performance.

Landscape evolution modeling (see, e.g., reviews by Bishop, 2007; Codilean et al., 2006; Coulthard, 2001; Martin & Church, 2004; Pazzaglia, 2003; Pelletier, 2013; Temme et al., 2013; Valters, 2016; Willgoose, 2005; Willgoose & Hancock, 2011) commonly couples together models for surface hydrology, hillslope sediment transport, and erosion by river channels. We call the specific rules used for each of these components *process laws*. Changes in the process laws (e.g., linear vs. nonlinear rule for hillslope sediment flux) represent changes in model structure. We refer to all elaborations of model structure, whether it be added parameters, new state variables, or new process laws, as the introduction of new *elements of complexity*.

For this study, we consider a set of models that represents a systematic sampling of the model space created by 12 binary choices in model construction. These choices fall into four categories: hillslope processes, channel processes, surface hydrology, and representation of geologic materials. Barnhart et al. (2020b, their section 4) reviews the basis for each of these permutations in the context of the study site. Each model structure requires between 2 and 10 input parameters (supporting information Table S1), reasonable ranges for which are discussed in Barnhart et al. (2020b, their section S5) and in a second companion manuscript (Barnhart et al., 2020c). The results of our sensitivity analysis are based on the sensitivity of each parameter with respect to the objective function. Based on a Method of Morris sensitivity analysis presented in Barnhart et al. (2020b), some model parameters were set constant for calibration (section 3.6).

Model analysis has previously been used in quantitative Earth surface research to infer timing of fault motion using 1-D models of scarp evolution (Andrews & Hanks, 1985; Andrews & Bucknam, 1987; Hanks, 2000; Pelletier et al., 2006), infer the form of governing equations most consistent with transient river long profile evolution (Attal et al., 2011; Hobley et al., 2011; Gran et al., 2013; Loget et al., 2006; Tomkin et al., 2003; van der Beek & Bishhop, 2003; Valla et al., 2010), and test alternative models of soil production and transport (Herman & Braun, 2006; Petit et al., 2009; Pelletier et al., 2011; Roering, 2008). In contrast, efforts to calibrate and validate models that couple hillslope and channel processes have been more limited, and generally focus on just one model (Gray et al., 2018; Hancock & Willgoose, 2001; Hancock et al., 2010; Willgoose et al., 2003; Ziliani et al., 2013). Studies that compare alternative coupled models are rare, and are typically limited to the comparison of just two model structures (Hancock et al., 2010). In this contribution, we advance the practice of model testing in quantitative geomorphology by systematically evaluating 37 alternative landscape evolution models at a field site with a well-constrained base level lowering history and access to high-resolution topography. This approach permits us to formalize a method of hypothesis testing in a way that is suited to our current understanding of landscape evolution theory.

2. Summary of Model Set and Implementation

The geomorphic basis for the 37 candidate models and the description of their governing equations presented in Barnhart et al. (2020b, their section S1). Here we briefly summarize the models and the reasons for choosing them.

2.1. Model Set

A summary of the twelve binary choices used to define the model set and the general form of the governing equations is provided below. From those twelve choices, we selected a small subset of the potential permutations and combinations (of which there are $2^{12} = 4,096$—too many for model calibration and validation). Barnhart et al. (2020b, their section 4.1) describes the basis upon which the full set of 4,096 possible models was reduced to 36. In this work we added one additional model, BasicChRtTh, to the set of 36 considered by Barnhart et al. (2020b), as described below.

The **default** and *nondefault* options for each binary choice are as follows (where the two-letter code labels the nondefault option):

1. Hillslope sediment flux as a **linear** or *nonlinear* function of local slope (Ch)

2. **Deterministic** or *stochastic* surface water hydrology (St)
3. **Uniform** or *variable source area* (VSA) runoff (Vs)
4. Channel erosion rule uses a **fixed** or *variable* exponent on drainage area. (Vm)
5. **No minimum threshold** or *a threshold* must be exceeded for channel erosion to occur. (Th)
6. **Stream power** or *shear stress* slope and area exponents. (Ss)
7. **Constant** or *depth dependent* channel erosion threshold. (Dd)
8. **Detachment-limited** or *entrainment-deposition* formulation for channel erosion. (Hy)
9. **Uniform** or *fine and coarse* fluvial sediment. (Fi)
10. **No separate soil layer** or *explicit soil layer*. (Sa)
11. **Homogeneous lithology** or *two lithologies*. (Rt)
12. **Constant climate** or *time-variable paleoclimate*. (Cv)

Each permutation was assigned a two-letter code, which is listed in parentheses. For example, the model using all default options is called Basic, whereas a model that considers two lithologies ("Rt") and an explicit soil layer ("Sa") is BasicRtSa. Each model also has a three-digit hexadecimal code (see Barnhart et al., 2020b, their Table 1, for detailed explanation). Table 1 lists all considered models. For description of the model governing equations, see Barnhart et al. (2020b, their section S1). Table S1 lists parameter symbols, names, and ranges. Table S2 lists parameters that freely varied in calibration for each model.

2.2. Implementation

The 37 models used here were developed using the Landlab Toolkit (Barnhart et al., 2020; Hobley et al., 2017) and are available in the `terrainbento` python package (Barnhart, Glade et al., 2019). It is important to note that Landlab is not, itself, a model. Instead it is a python package containing a gridding engine and a number of model components, each treating one physical process (e.g., hillslope sediment flux by linear diffusion). A benefit of developing Landlab-built models is that the difference between alternative models can be isolated.

We implement our models at a horizontal grid cell spacing of 7.3 m (24 ft) and a time step of 10 years (1 year for stochastic precipitation models). Barnhart et al. (2020b, their section 6) connects our discretization choices with geomorphic process and numerical constraints. Each simulation runs from 13 ka, the time of last deglaciation, to the present day.

Models are initialized with topography that represents the postglacial topographic surface. The Method of Morris sensitivity analysis of Barnhart et al. (2020b) indicates that, in terms of the objective function (which is described further in section 3.3), the models are not sensitive to the choice of initial conditions. As such, we use the case of 7% channel etching and no change in the upper watershed, as discussed in Barnhart et al. (2020b, their section 8.2.1).

3. Calibration Approach

The goal of model calibration is to find the set of model input parameters that minimizes the objective function (e.g., B. Adams et al., 2017a, 2017b; Hill & Tiedeman, 2007; Tarantola, 1987; Tarantola & Valette, 1982). This is equivalent to finding the *global minimum* of the objective function surface in a parameter hyperspace with a number of dimensions equal to the number of estimated parameters. This minimum represents the "best" parameter value set, conditional on the definition of the objective function (section 3.3).

There are two major classes of calibration algorithms. *Gradient-based methods* use the gradient of the objective function surface to search for a minimum (see B. Adams et al., 2017a, 2017b, their Chapter 6 for additional background on optimization). In contrast, *global methods* take a sampling approach to identify a global minimum. A subtype of gradient-based methods that we employ is the *nonlinear least squares* method, a type of optimization algorithm that exploits the mathematical properties of a sum of squares objective function like the one we define in section 3.3. It is also important to note that some calibration algorithms use only *complex model evaluations* (i.e., evaluations of the full model), while others use complex model evaluations to construct a *statistical surrogate model*. Surrogate models use complex model evaluations

Table 1
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.

sampled from parameter space to approximate the objective function surface used by the calibration algorithm to find the optimal parameter set.

Simulations at 7.3 m resolution take approximately 30 min or more on the computing cluster we employed. As such, a computationally frugal calibration method is necessary for this application. We found that successful calibration required a hybrid calibration approach, one that uses a surrogate-based, global method followed by a complex model, gradient-based, local method. We used Sandia National Lab's Dakota package

to manage model analysis in this work (B. Adams et al., 2017a, 2017b). The surrogate-based global method is constructed using Efficient Global Optimization (EGO, Jones et al., 1998). In the EGO method, model evaluations are iteratively made to create and refine a statistical (Gaussian process) surrogate model of the objective function. We use EGO to find the region of the global minimum. We then refine the estimate of the global minimum using a second optimization method, NL2SOL, which is a gradient-based method that is well suited to least squares problems with large residuals (Dennis et al., 1981). NL2SOL is started from the best point in parameter space found by EGO. Examination of selected models suggested that this approach is able to identify optimal parameter sets that meet the success criteria defined in section 3.2 below.

3.1. Nonlinear Least Squares Optimization

Nonlinear least squares methods estimate the vector of optimal parameter values, $\boldsymbol{\beta}$, that minimize the sum of squares of the residuals between observations, $\mathbf{y}$, and simulated equivalents, $\mathbf{y'}(\boldsymbol{\beta})$, as weighted by a weight matrix, $\mathbf{w}$ (B. Adams et al., 2017b; Dennis Jr & Schnabel, 1996; Hill & Tiedeman, 2007). The general form of the objective function, F_{obj}, is defined as

$$F_{\text{obj}} = \left[\mathbf{y} - \mathbf{y'}(\boldsymbol{\beta})\right]^T \mathbf{w} \left[\mathbf{y} - \mathbf{y'}(\boldsymbol{\beta})\right] . \tag{1}$$

This form of objective function is commonly called the $L2$ norm as it contains squared residuals.

Optimal parameter values are estimated by iterative refinement of the values in $\boldsymbol{\beta}$ such that F_{obj} is minimized. Estimated parameter uncertainty is obtained from the parameter variance-covariance matrix $\mathbf{V}(\boldsymbol{\beta})$, given by

$$\mathbf{V}(\boldsymbol{\beta}) = s^2 \mathbf{J}^T \mathbf{w} \mathbf{J} \tag{2}$$

where $\mathbf{J}$ is the *sensitivity matrix* (also called the Jacobian), which describes the partial derivatives of $\mathbf{y} - \mathbf{y'}(\boldsymbol{\beta})$ with respect to each calibrated parameter in $\boldsymbol{\beta}$, and s^2 is the calculated error variance. The latter is defined using the bias-corrected version as (Hill & Tiedeman, 2007, their equation (6.1))

$$s^2 = \frac{F_{\text{obj}}}{N_d + N_{pr} - N_p} . \tag{3}$$

Here N_d is the number of observations, N_{pr} is the number of prior information values, and N_p is the number of calibrated parameters for a given model. For our application $N_d = 20$, one for each of our landscape patches, $N_{pr} = 0$, and N_p ranges from 2 to 7. It is required that $\mathbf{J}$ is not rank deficient (B. Adams et al., 2017b; Gill et al., 1981).

We initially hypothesized that it would be possible to calibrate the suite of models using only the Gauss-Newton algorithm, a computationally frugal gradient-based algorithm (Hill & Tiedeman, 2007, pp. 68 and 77). However, preliminary trials with the Gauss-Newton algorithm revealed many local minima, which meant that the gradient-based method alone was not suitable (section 4.2).

3.2. Definition of Multimodel Calibration Success

Unless it is possible to prove that the objective function is convex in the considered portion of parameter space, it is not possible to know *a priori* whether *local minima* in the objective function surface are present. When there is an analytical expression for the second derivative of the objective function, this claim can be proven. However, in the context of landscape evolution models and the objective function used in this study, no analytical expression exists that relates the input parameter values to the objective function. Thus it is not possible to prove or disprove convexity *a priori*. When convexity is not guaranteed, it is not possible to know whether the parameter set produced by an optimization algorithm represents the *global minimum* or one of an unknown number of possible *local minima*. To address this issue in our multimodel calibration effort, we created a set of criteria for determining whether a calibration is successful. Similar criteria have been proposed and used by McKenna and Poeter (1995), Poeter and McKenna (1995), and Foglia et al. (2013).

Criterion 1. Simulated terrain evolution should make sense given the process representation within the specified model. For example, a model that uses a linear relation between hillslope sediment flux and slope is expected to simulate plateau edges that are smoother than the real ones. Simulations that contradict basic geomorphic principles such as this would be suspect and therefore rejected.

Criterion 2. Model ranking based on an objective function should not be in conflict with model ranking by experts based on simulated equivalents.

Criterion 3. The best fit parameters obtained from the calibration should not fall on the boundaries of the parameter space domain. If the best fit value of a particular parameter ends up at the high or low extreme of its range, then we cannot assess whether this result represents a global minimum. If calibration were to identify a best fit location at the edge of the parameter range, it could indicate a model deficiency: the model is simply not correctly representing processes related to that parameter. Alternatively, it could indicate that a model is adjusting such that it can recover a simpler model through parameter choice (for example, a threshold value might calibrate to 0, which indicates that the threshold term does not add explanatory power).

A practical exception to this third criterion is the case of relatively unimportant parameters (such as the hillslope diffusivity, D) that have little impact on the objective function. In these cases we consider values near the edge of the parameter range as permissible.

Criterion 4. Any multielement model that has the Basic model as a special case should have a best fit objective function score that is less than or equal to that of the Basic model. In these cases, it is expected that adding a degree of freedom should improve model performance. For example, model BasicVm is identical to Basic in all respects except that it treats the drainage area exponent m as a calibration parameter, rather than fixing it at 1/2 (see Barnhart et al., 2019). Because $m = 1/2$ falls within the assigned parameter range, it is possible for BasicVm to exactly mimic Basic. For models in this category, their best fit objective function score should be at least as good (i.e., at least as small) as the score for Basic. We interpret failures of such a model to outperform Basic to indicate that the calibration is stuck in a local minimum in objective function space. If the parameter space does not allow a model to mimic Basic, this criterion is not applied. For example, because model BasicSa must produce regolith in order to move it, and erosion rates at the site are high (~4 m/ka), soil production would need to be unrealistically high to reproduce the dynamics of the Basic model, which requires no weathering to generate hillslope sediment flux.

More generally, any multielement model whose range includes one or more simpler models should outperform the simpler model(s). For example, model BasicRtTh, which uses an erosion threshold ("Th") and treats rock and till separately ("Rt"), should outperform both BasicRt and BasicTh.

Criterion 5. A model calibration should be able to complete in a reasonable amount of time. "Reasonable completion time" was considered to be a single model evaluation (that is running a simulation for one set of parameters) could not take more than 24 hr on one core of the University of Colorado Summit heterogeneous supercomputing cluster. Calibration of each model required many model evaluations. The 24 hr limit is due to the wall time limits for entry into the standard queue on Summit. We did tests to determine whether relaxing this limit from 24 hr to 7 days resulted in increased completion, and it did not. Seven days is the longest possible job time on Summit, so increasing job length beyond this limit was not possible.

Exploratory calibration using an objective function composed of weighted topographic metrics (such as the hypsometric integral) failed to meet Criteria 1–5. We addressed this by assessing and refining the calibration algorithm and sum of squares objective function elements until we found reasonable results. This is discussed further in section 4.2.

Three models failed to complete individual simulations within the constraints of Criterion 5: BasicCh (040), BasicHySa (410), and BasicChSa (440). Model BasicHySa is the model in which numerical stability is most sensitive to model time step, and parameter values are known to significantly affect the maximum stable timestep in this class of model (Shobe et al., 2017). Examination of model log files indicated that under certain parameter combinations, very small timestep size was required. This was not possible given walltime constraints. Models BasicCh and BasicChSa have an internal routine that reduces sub–time step duration when needed to ensure numerical stability. When slopes are especially steep, the solution routine can demand very small internal timesteps, which in turn leads to prohibitively long run times. Because the most successful models turned out to be those that incorporate separate rock and till lithologies, and the three uncalibrated models lack this feature, it is likely that they would not have scored well had their calibrations completed.

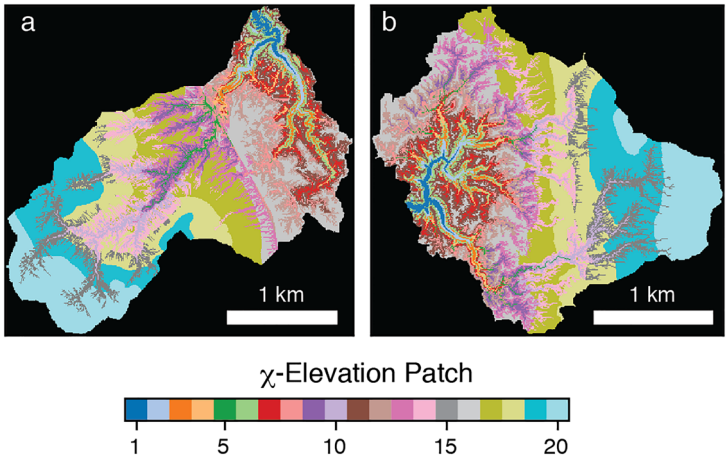

Figure 2. Maps of χ elevation categories for the calibration domain (a) and the validation domain (b).

3.3. Objective Function Definition

Our definition of model performance uses an objective function based on a direct cell-by-cell comparison of modeled and observed DEMs (see Barnhart et al., 2020b, their section 6 for full definition). Use of the nonlinear least squares optimization method described in section 3 requires fewer simulated equivalents than the $\sim 10^5$ model grid cells, but more than the number of estimated parameters. To accomplish this we constructed an objective function composed of residual scores from a set of 20 patches that broadly represent characteristic landform features (Figure 2). Patches were defined based on two criteria: elevation and the χ index value (Perron & Royden, 2013), which is a topographical property that can be related to equilibrium channel longitudinal profiles (Harkins et al., 2007). We found that χ was useful to distinguish between similar process domains that are not necessarily spatially adjacent.

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

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

P_j is calculated as

$$P_j = \sqrt{\sum_{i=1}^{N_j} w_i \left(\eta_i^{\text{obs}} - \eta_i^{\text{sim}}\right)^2}, \tag{5}$$

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. The weight factor for cell i in patch j is given by

$$w_i = \frac{1}{\sigma_i^2 N_j}, \tag{6}$$

where N_j is the total number of grid cells in patch j, and σ_i is the initial condition elevation uncertainty. We include the number of cells in the patch in order to weight some patches more than others. Patches in the lower reaches of the watershed are designed to have fewer cells and thus be emphasized more in the calibration. Barnhart et al. (2020b, their Figure 4) describes the definition of spatially variable σ. The objective function calculations were implemented using the `umami` python package (Barnhart, Hutton, et al., 2019).

3.4. Experimental Design

For each of the 37 alternative models, we attempted an independent calibration to estimate each model's parameters. All calibrations were constrained to predetermined parameter ranges. The number of parameters permitted to freely vary in calibration was determined using the Method of Morris sensitivity analysis results described in the first companion paper (Barnhart et al., 2020b). The permitted range for each parameter was based on the second companion paper (Barnhart et al., 2020c) and is given in Table S1. A list of

all parameters varied in calibration is given by Table S2. Three of the models failed to meet Criterion 5 (reasonable completion time), and were therefore not considered further.

As an independent check on the calibration, we ran the 34 calibrated models on the five alternative validation domain postglacial topographies, and calculated the objective function for each (Table S38). Comparing the calibration and validation results allows us to identify whether the relative rank-ordering of models changes when the models are run in a similar watershed.

3.5. Metrics of Model Evaluation

We compared the calibrated models with one another based on the minimum objective function value for each model. Objective functions are inherently statistical quantities, and when objective function values from different models are compared, this statistical character needs to be considered. This is accomplished by comparing confidence regions around objective function minima. Following Hill and Tiedeman (2007, p. 178, equation (8.14), p. 178, their Table 8.2), the $(1-\alpha)100\%$ confidence interval CI_α can be calculated as,

$$CI_\alpha = F_{\text{obj}} \pm s^2 c_{(1-\alpha)100} \tag{7}$$

where s^2 is as defined in equation (3) and $c_{(1-\alpha)100}$ is a critical value.

Confidence intervals based on critical values from the Student t distribution are sometimes considered to be too small, while those with critical values from the F distribution are too large (Hill & Tiedeman, 2007). Thus we calculate and report confidence intervals based on critical values from both the Student t and F distributions. The symmetric confidence intervals implied by equation (7) are valid for symmetric objective functions. In section 4.1 we explore the objective function of model 000 Basic and find that the objective function is weakly asymmetric.

Models with more calibration parameters are expected to perform better than models with fewer calibration parameters, all else equal, simply because there are more fitting parameters. The new parameters can only be considered to significantly improve the model if the improvement is "sufficient"; in other words, if the improvement in objective function is more than would be expected from the addition of an extra fitting parameter. Several model comparison metrics have been developed that combine the objective function with penalty terms based on the number of calibration parameters. With these metrics, the definition of "enough" is that the improved model fit needs to overcome the penalty incurred by the added parameters.

Following the recommendations of Burnham and Anderson (2003, p. 66), we use the corrected Akaike Information Criterion AIC_c (Akaike, 1973, 1974; Poeter & Hill, 2007; Sugiura, 1978):

$$\text{AIC}_c = F'_{\text{obj}} + 2N_p + \frac{2N_p(N_p+1)}{N_d + N_{pr} - N_p - 1}\ . \tag{8}$$

where F'_{obj}, the maximum likelihood objective function, is defined after Hill and Tiedeman (2007, Appendix A)

$$F'_{\text{obj}} = (N_d + N_{pr}) \ln 2\pi - \ln|\omega| + F_{\text{obj}}\ . \tag{9}$$

Here $|\omega|$ is the determinant of the weight matrix. As we weight each patch equally (section 3.3), in our application ω is a vector of ones of length N_d and $\ln|\omega| = 0$.

Thus, a model with more calibrated parameters will be penalized by having a higher value of AIC_c, all else equal. For example, the Basic model has two calibrated parameters and for it: $\text{AIC}_c = F'_{\text{obj}} + 4 + 12/17 = F'_{\text{obj}} + 4.7$. In contrast, model BasicSt has five calibration parameters. For it $\text{AIC}_c = F'_{\text{obj}} + 10 + 60/14 = F'_{\text{obj}} + 14.3$. The BasicSt model must improve the value of F'_{obj} by 9.6 in order to be comparable to the Basic model. For our application, model ranking based on AIC_c and F_{obj} are nearly identical. In plots presented here we show F_{obj}.

3.6. Parameters Set Constant in Calibration

A Method of Morris sensitivity analysis by Barnhart et al. (2020b) revealed that many parameters exert little influence on the objective function. By holding these parameters constant in the calibration process, we can significantly reduce the computational and analytical complexity of calibration. Table S2 lists the parameters varied during calibration.

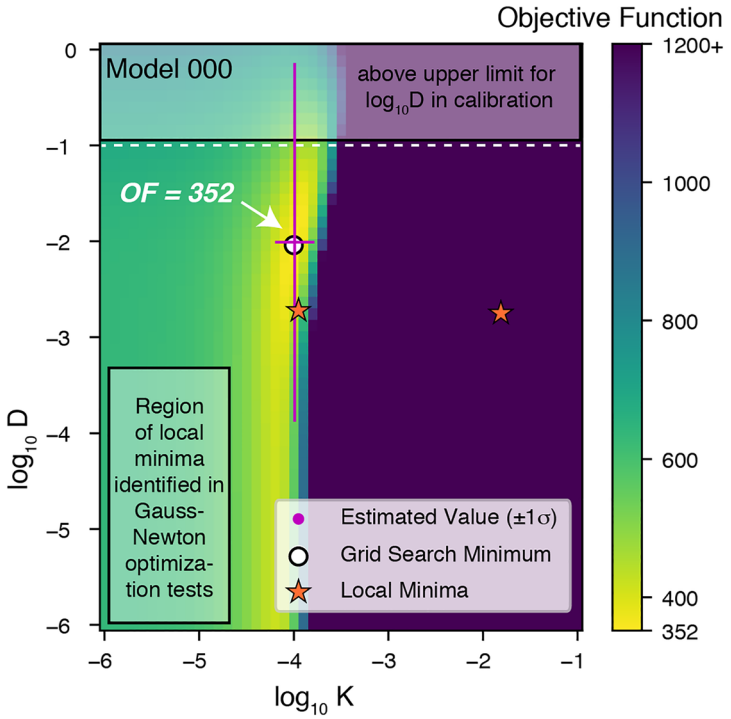

Figure 3. Objective function surface for model Basic (000) characterized by a grid search of size 51 × 51 (a total of 2,601 model evaluations). The dashed white line indicates the upper limit of the diffusivity parameter range considered in sensitivity analysis and calibration. Orange stars identify local minima. Magenta lines show optimum point and ±1σ bounds identified by hybrid optimization (see text).

The hillslope transport efficiency coefficient, D, appears in all models. The log-transformed coefficient, $\log_{10}D$, rarely emerged among the most influential parameters. Nonetheless, we treat it explicitly in calibration because it is the primary—and for many models, only—parameter to describe the rate of downslope motion on hillslopes.

The random seed used in stochastic precipitation models was treated as a parameter in the sensitivity analysis in order to test the influence of the particular random sequences used. The seed value had uniformly low influence. This finding implies that the differences between one random sequence and another (both drawn from the same underlying distribution) have little impact on model output. The random seed was held constant in calibration. Six models with stochastic precipitation show little sensitivity to either the precipitation distribution shape factor, c, or to the number of sub–time steps used in the numerical algorithm, n_{ts}. As such, both parameters are held constant in model calibration.

Models with a dynamic soil layer (Sa) are relatively insensitive to the characteristic soil thickness, H_0. This parameter represents the length scale over which weathering rate declines. For calibration it was set to 0.5 m. Models that include variable source area (VSA) hydrology, as well as those with a dynamic soil layer, use the parameter H_{init}, the initial soil thickness. In both cases, the models show little sensitivity to the parameter. In calibration, it is set to 1.5 m based on the observed thickness of soils at the site (Barnhart et al., 2020b).

Models that include VSA hydrology also specify a recharge rate, R_m. This parameter is one of three that control subsurface flow capacity, with the others being soil thickness (H_{init}, or dynamic if applicable) and saturated hydraulic conductivity, K_{sat}. Together these three parameters effectively form a single lumped parameter, and thus it is only necessary to calibrate one of them. The recharge R_m is held constant at 0.5 m/yr (roughly half the site's mean annual precipitation) while K_{sat} is retained as a calibration parameter. The models with two lithologies (Rt) are generally insensitive to the width of the contact zone between glacial sediments and bedrock, W_c. For calibration, it is held fixed at 1 m.

4. Results

As it is not common to formally calibrate landscape evolution models, we first describe the results of detailed interrogation of the objective function for model 000 Basic. After our initial calibration attempt we determined that a global calibration method was necessary and made a second calibration attempt. This resulted in 34 of the 37 considered models successfully calibrating (as defined by our criteria for calibration success).

4.1. Model Basic Objective Function Surface

Figure 3 shows the objective function surface for two-parameter model 000 Basic as determined by a simple grid search. This model is one of only two models in the collection that have only two calibration parameters. Because the parameter space is only two-dimensional, we can easily visualize the objective function surface. Exploring this model's objective function is also useful because model Basic contains two geomorphic transport parameters that are present (in some form) in all of the models: D, a parameter controlling soil gravitational transport, and K, a parameter controlling the efficiency of water erosion.

The Basic objective function surface shows three primary domains (Figure 3). First, in light green and on the left, is a flat region with objective function values that are relatively low, but not as low as the global minimum. This region corresponds to model runs in which little or no erosion occurred. This feature of the objective function makes sense given that the overall shape of the postglacial topography and the modern topography are similar; models with little erosion get a moderate score that reflects their "success" in not eroding the preserved plateau remnants. Second, in purple and on the right, is a region with very high objective function values. In this region, too much erosion occurred due to high values of K, the parameter that controls the ability of streams to incise. Finally, between these two regions lies a narrow band of lower objective function values (in yellow on Figure 3). For values of D less than about 10^{-2} m^2/yr, the location of this trough is only influenced by the value of K. For higher values of D, we see that the orientation of the trough is influenced by both D and K. Adjacent to the grid search minimum point (shown with a white dot) is a region with similarly low objective function values. We examined the objective function for local minima (at the scale of our grid search) and found two, including one near the global minimum (orange stars in Figure 3).

Examining the Basic model objective function surface, we can conclude that it is non-Gaussian. A Gaussian objective function is one of the assumptions underlying many nonlinear least squares calibration methods. The actual objective function surface shows an extensive region of parameter space that is relatively flat, and therefore challenging for gradient-based optimization methods. Careful choice of convergence criteria and algorithm step size are likely to be important for successfully applying a gradient-based method like Gauss-Newton. The discrete grid values of Figure 3 reveal one local minimum near the global minimum. We expect that there are additional local minima not observed at the resolution of the grid search because multiple exploratory calibrations found that the Gauss-Newton method converged in the lower left corner of Figure 3.

4.2. Initial Calibration Attempt with Gradient-Based Algorithm

Initial Gauss-Newton model runs with model Basic (000) successfully found the approximate location of the global minimum shown in Figure 3. In initial calibrations of Basic and in follow-on trial calibrations with other models, we found that many models produced calibration results with unrealistically low K and D values (shaded region in lower left of Figure 3), resulting in little to no erosion. In order to understand this issue, we performed a high-resolution series of parameter studies on the Basic model. Figure 4 shows two transects approximately through the objective function minimum determined based on an early grid search (Figure 3). The transects also demonstrate the presence of many small local minima. In light of this finding, we decided to pursue an alternative method for calibration that is more robust with respect to local minima.

4.3. Calibration of Models With a Hybrid Method

After attempting calibration with the Gauss-Newton method, we explored calibration with a number of non-surrogate global methods available in Dakota, including Adaptive Mesh as well as Direct (see B. Adams et al., 2017a, for details and references related to these methods). We found a workable calibration methodology that first uses the EGO surrogate-based global method (Jones et al., 1998) and then refines the search using the NL2SOL method (Dennis et al., 1981). We settled on this method because it was the first calibration method that we tried that provided results that met the criteria for calibration success of section 3.2. First,

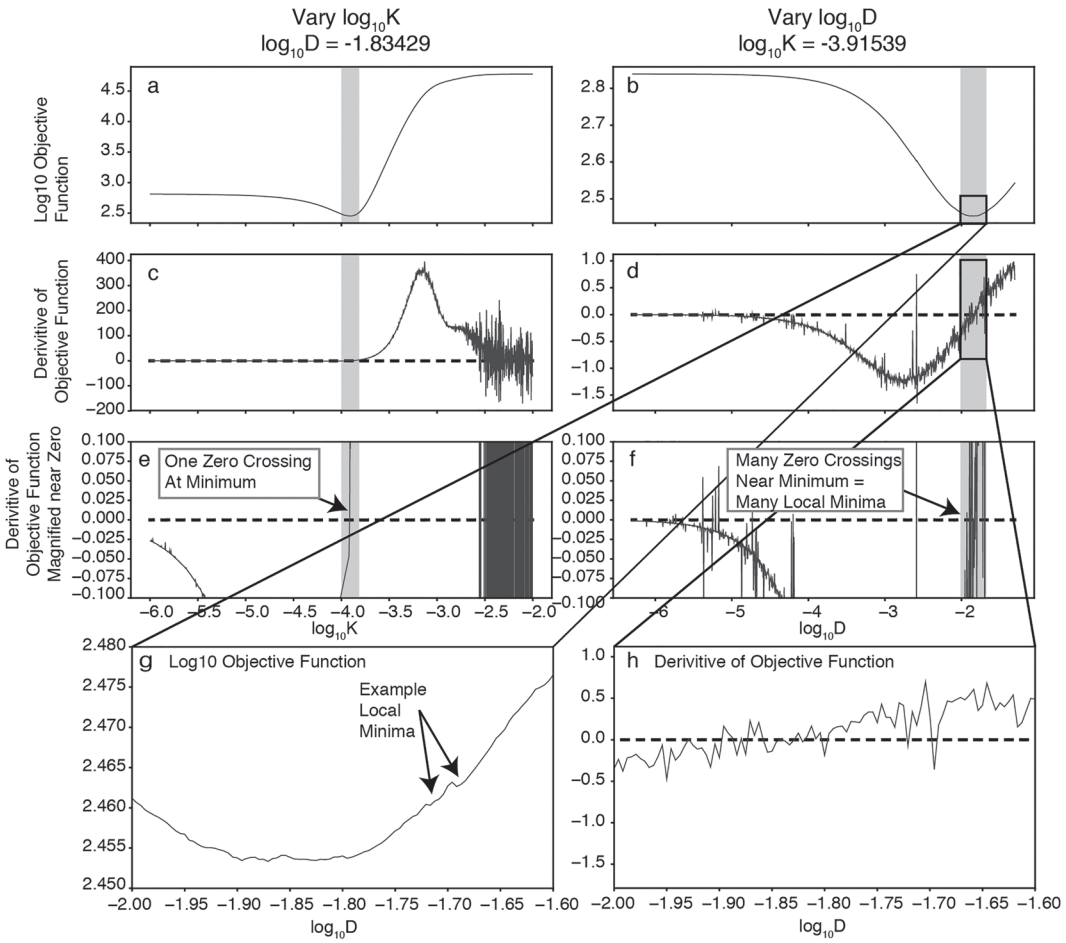

Figure 4. Results of a transect-based parameter study of model 000 Basic. The top three panels of the left-hand column show a parameter study in which $\log_{10}K$ was varied and $\log_{10}D$ was fixed, and the top three panels of the right-hand column show a parameter study in which D was varied and $\log_{10}K$ was fixed. The gray bars highlight the region of the global minimum. The top two panels (a and b) show the $\log_{10}$ of the objective function, while the second row of panels (c and d) show the first derivative of the objective function. The third row of panels (e and f) show a zoomed-in version of panels c and d. Finally, panels g and h show magnified portions of panels b and d. As panel g shows, the objective function is bumpy, and this is manifested in multiple zero crossings in panel h. While these only show bumps in 1-D, they are an indication that local minima in 2-D exist.

the efficient global method finds the approximate location of the global minimum, and then the parameter values are refined by a gradient-based method.

Figure 5 presents maps of modern minus end-of-model run topography (here called the *topographic residual*) for all calibrated models, and Figure 6 shows the objective function values with their uncertainty bounds. There is a clear break between the performance of the first eight models—which all differentiate between rock and till (Rt models)—and the remaining models.

One rock-till model performs notably worse than the rest. It includes explicit treatment of soil, and is discussed in section 5.1.4. All model fit metrics are presented in Table S3, and a hillshade of each model's end-of-run topography is given in Figures S2–S35. Assessing model rank using the AIC_c, which includes penalty terms for models with more parameters, does not substantially impact the relative ranking of models. Using the variance of the least squares estimator, we obtain estimates of each model parameter (Tables S4–S37).

Using a grid search, we verified that the hybrid EGO-NL2SOL method found the correct region of the objective function minimum for models Basic and BasicRt. We also verified that, as expected, the EGO-NL2SOL method always improves upon the EGO method alone (Figure S1). Given that we are unable to demonstrate convexity for our models, and run times prohibit a comprehensive grid search for any but the simplest

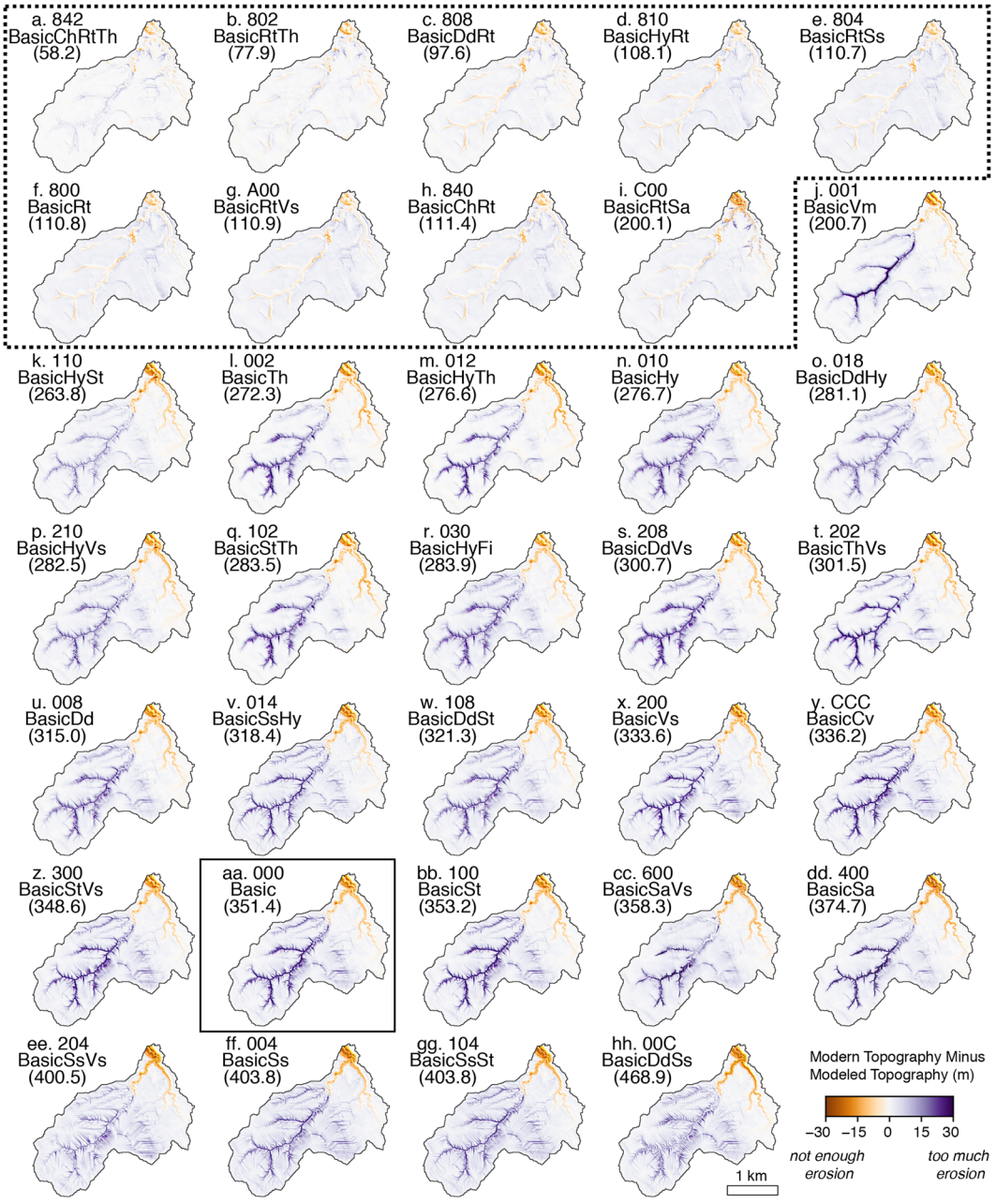

Figure 5. Maps of model fit (observed minus simulated modern topography) for all 34 successfully calibrated models ordered from best (a) to worst (hh). The dashed line outlines models with the Rt element, and the solid line identifies model Basic. Value given in parentheses is the model's best fit objective function score.

models, we must accept that an unknown number of subsequent solutions are local minima. However, we designed the success criteria carefully with this reality in mind, and expect the effect to be minor. The combination of a surrogate-based global method (e.g., EGO) and a gradient-based method (e.g., NL2SOL) is a promising strategy for calibrating landscape evolution models.

Several models completed calibration with estimated parameter values on the edge of the parameter space. This typically occurred for one of three reasons:

1. A calibrated value indicated that the model was trying to recover a simpler option. For example, model 840 tried to recover the dynamics of model 800 by setting the critical slope S_c to the highest possible value.

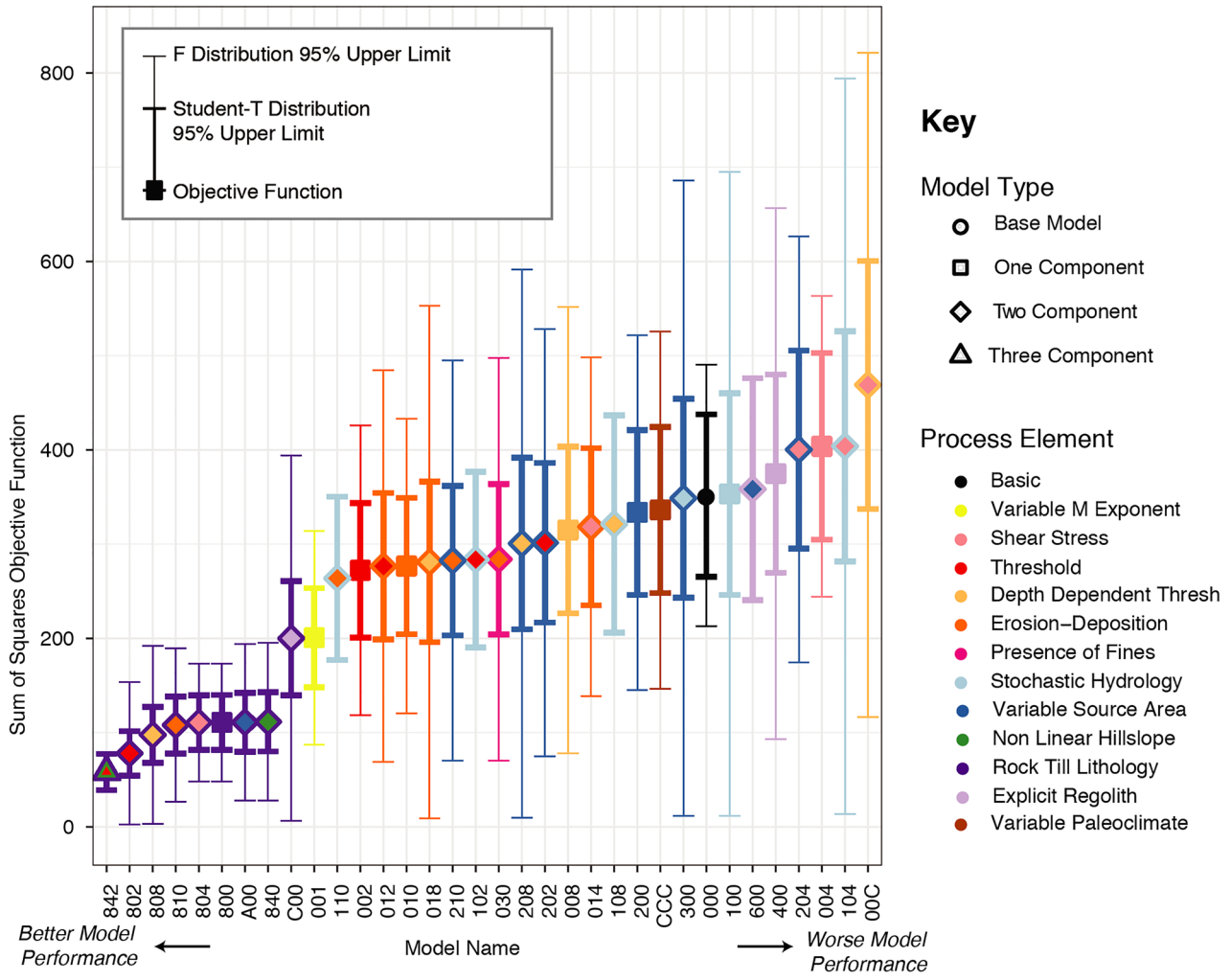

Figure 6. Model fit obtained in calibration measured using a sum of squares objective function (section 3.3). Models are ranked based on calibrated value from left to right. Models may have a higher (worse) objective function value than Basic (black dot) for two reasons: (a) a model has a similar objective function value to Basic but was penalized for having more parameters, or (b) the model cannot fully recapture the functional form of Basic through parameter values (see section 3.2, Criterion 4). Two confidence intervals are shown for each model: one that likely exaggerates uncertainty (*F* distribution, thin lines), and one that underestimates uncertainty (Student *t*, thick lines).

2. The calibrated value of the hillslope diffusivity D was at the highest possible value. We do not consider this to indicate calibration failure because the objective function shows low sensitivity to D (Barnhart et al., 2020b). We included it in the calibration as it is often the only parameter that influences hillslope sediment flux.
3. The calibrated value of the maximum soil production rate P_0 was at the highest possible value, and the soil depth-dependent hillslope sediment flux decay depth was estimated at its lowest possible value. This result is discussed further in section 5.1.4.

5. Discussion

5.1. Which Model Elements Improve Performance?

The calibration results overwhelmingly support the conclusion that models for the Franks Creek watershed that differentiate between bedrock and glacial till perform better than models that do not. These models all include the Rt component of Table 1. This is evident in the group of best performing models, which all have purple outlines in Figure 6 indicating rock-till differentiation. The effect of including the Rt element on the results is that the erodibility coefficient for bedrock is allowed to be smaller than that for till. This permits the incision of the channel network into the till plateau without extensive incision of the upper watershed

(Figure 5). The addition of this element of complexity reduces the objective function value by a factor of three, from 351.4 to 110.8.

To illustrate this conclusion, we contrast results from model Basic (000) and model BasicRt (800), which differ only in the inclusion of the Rt component (Figures 5f and 5aa). The objective function was designed to put more weight on erosion in the lower part of the drainage network and adjacent plateau. The calibrated model Basic over-incises in the upper parts of the watershed in order to enable some incision in the lower part of the watershed. This results in modeled modern topography that has not incised enough in the lower part of the watershed but has incised too much in the upper part of the watershed. In contrast, model BasicRt has two values for erodibility, and is thus able to incise more in the lower part of the watershed while not incising excessively in the upper, bedrock-underlain portions of the watershed. There are still flaws in model BasicRt's performance. For example, the valley it incises into the till plateau is too narrow. The overly narrow valleys in both models reflect the use of a simple linear diffusion law for hillslope transport, rather than a nonlinear law with a specified slope threshold (see Barnhart, Glade, et al., 2019). Additionally, model BasicRt erodes more than it should in areas between channels, over much of the watershed.

Among the eight two-element Rt models, the most successful are those two that include an erosion threshold, BasicRtTh (802) and BasicDdRt (808) (Figures 5b and 5c). Model BasicRtTh includes two additional parameters that model BasicRt lacks: an erosion threshold for glacial sediments and an erosion threshold for bedrock. In model BasicDdRt, which allows the erosion threshold to increase with incision depth, it is the rate of change of threshold value with incision depth that varies between rock and glacial sediments. The addition of the erosion threshold permits the calibrated models to incise the main channels but not over-erode away from the main channels. The problem of insufficient erosion along the side slopes of lower Franks Creek remains, but as in the case of model Basic, this is to be expected because neither BasicRt or BasicRtTh has a mechanism to create planar hillslopes.

Given the preliminary success of model BasicRtTh and the anticipated improvement the nonlinear hillslope component would provide in better capturing the planar slopes adjacent to Franks Creek, we created an additional, three-element model, BasicChRtTh (842). This model retains the rock-till map and use of erosion thresholds, and it also adds a nonlinear (Taylor series) model of downslope soil motion (see Barnhart, Glade, et al., 2019, for details on the formulation of this model). The nonlinear law has the property that it tends to create planar side slopes with a gradient close to a specified threshold gradient, S_c. This model performs the best out of all calibrated models (Figures 5a and 7d).

The importance of nonlinear hillslope sediment transport in calibration provides an apparent contrast with the results of Barnhart et al. (2020b), who found that the hillslope diffusivity D is not often an important parameter in sensitivity analysis. As discussed further in Barnhart et al. (2020b, their section 10.6) this seemingly contradictory finding points to the important distinction between an important *process* and an important *parameter*.

5.1.1. Nonlinearities in Model Structure

Model BasicChRtTh performs the best of all considered models. But BasicChRt—which combines a nonlinear hillslope law with a rock-till map, but unlike BasicChRtTh, lacks a threshold—does not perform substantially better than model BasicRt (they have nearly identical objective function values; Figures 5f and 5h). Moreover, the calibrated value of its threshold gradient (S_c) parameter is at the upper limit of the permitted parameter range. This is an indication that in calibration BasicRtCh is attempting to recover the diffusive end member presented by BasicRt (Table S33). Taken together this indicates that the Ch and Th model structure elements are interacting nonlinearly.

Figure 7 contrasts four calibrated models to explore this: BasicRt, BasicRtTh, BasicChRt, and BasicChRtTh. Comparison of the hillshade and topographic residual reveals where each model made advances. The addition of a threshold reduces erosion slightly over the entire domain (purple goes to white from Figures 7a to 7b). The addition of the nonlinear hillslope law widens the main valleys in the lower part of the basin (orange gets lighter in some locations from Figures 7b to 7d). Examining the areas adjacent to the channels in the upper portion of the watershed in BasicChRtTh indicates that these areas have eroded more than in BasicRtTh. This represents the nonlinear hillslopes responding to channel incision in the upper watershed, and it indicates why BasicChRt did not improve over BasicRt.

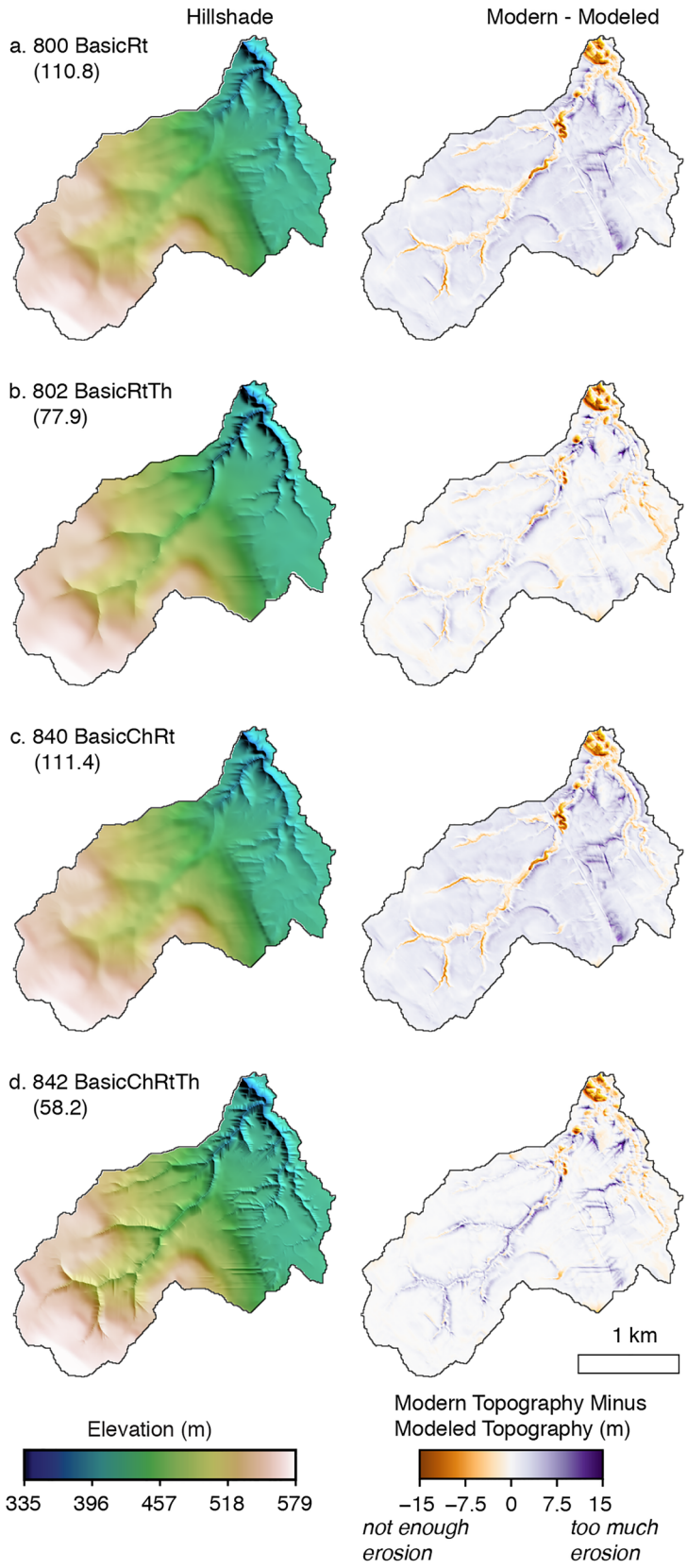

Figure 7. Comparison of hillshade (left column) and modeled topography (right column) for four models that include the rock-till distinction. (a) 800 BasicRt (b) 802 BasicRtTh, (c) 840 BasicChRt, (d) BasicChRtTh.

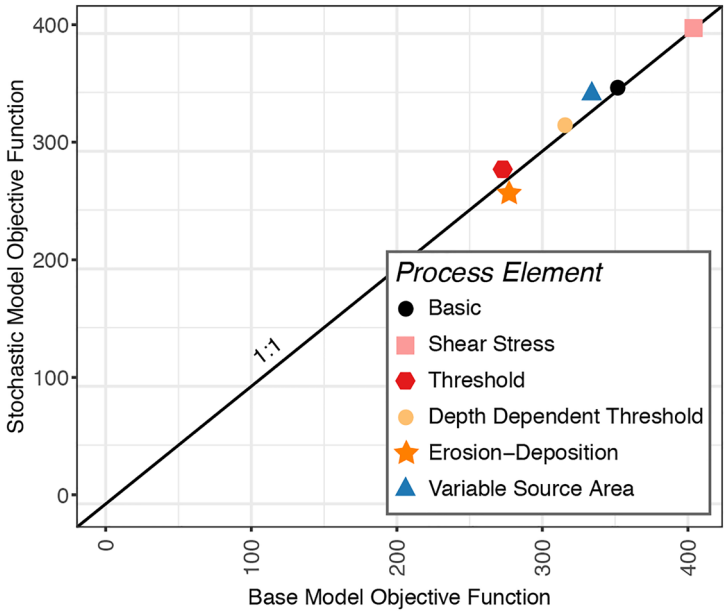

Figure 8. Comparison between base models and stochastic models for the six model pairs that include adding the St element. Colors indicate the nonstochastic model element. For example, the black dot indicates a comparison between Basic and BasicSt, and the red hexagon indicates a comparison between BasicTh and BasicStTh. A 1:1 line is shown for reference.

We conclude that unrealistically wide gullies in the upper watershed occur in model BasicChRtTh because stream reaches with small contributing area in the upper part of the watershed experience water erosion and then create steep slopes that result in increased hillslope sediment flux. This then results in too much erosion in the upper part of the watershed. This result points to potential value of letting hillslope parameters vary with lithology. Modifying the nonlinear hillslope component to allow for a spatially variable S_c, with different values assigned to bedrock and till domains, could lead to improved performance by allowing steeper valley side slopes in the bedrock portion of the watershed.

5.1.2. Coincidence of Drainage Area and Lithology

The best performing model that does not include rock and till is model BasicVm (001). The performance of this model reflects the coincidence between drainage area and lithology: the largest streams flow on the weaker till material. The process change present in model BasicVm is a variable drainage area exponent. In model Basic, the value for m is set at $1/2$, which means that, for a given channel slope, erosion rate is proportional to the square root of drainage area. This has the effect of more channel incision by water erosion in the downstream reaches of the major streams, where drainage area is larger. In the case of Franks Creek watershed, these downstream reaches happen to correspond to the areas with thick till and deep incision. The calibrated value of m in BasicVm is 0.86, which describes a faster downstream increase (relative to the $m = 1/2$ case in the Basic model) in the efficiency of water erosion with increasing drainage area. In other words, the calibration of model BasicVm uses a higher-than-expected value of m to compensate for the assumption of uniform lithology and allow the lower reaches of channels that cross the till plateau to incise more deeply. The modification of m is an example of obtaining the right results for the wrong reason (Beven, 1989; 2002; Grayson et al., 1992; Kirchner, 2006; Klemeš, 1986; Lane & Richards, 2001).

5.1.3. Extra Complexity With No Improved Performance

One might expect that stochastic hydrology models (St: 100, 102, 104, 108, 110, 300), which have three calibration parameters that control rainfall intensity and frequency, would be able to outperform their deterministic counterparts. Our suite of 37 models includes six pairs of stochastic/deterministic models—that is, models that are identical in every respect except that one is deterministic and the other is stochastic (Figure 8). Examining the relative performance of these pairs permits us to identify whether explicitly treating stochastic variability in runoff improves model results. We find that stochastic models do not calibrate any better than their deterministic counterparts, even given their additional calibration parameters. We

interpret this as an indication that explicitly treating the rainfall distribution does not provide additional explanatory power in this region. This result indicates that for our application, the use of an "effective" erodibility factor (and, for fluvial threshold models, a smoothed threshold) appropriately subsumes the effects of sequences of high and low runoff. This finding is consistent with previous analyses of how stochastic variability in streamflow impacts erosion and sediment transport (Lague et al., 2005; Molnar, 2001; Tucker & Bras, 2000; Tucker, 2004; Willgoose et al., 1991). A common thread in these analyses is that the influence offlow variability can be expressed analytically through a factor that includes an integral over all possible discharge values, thereby providing an expected rate of erosion or sediment transport when averaged over all possible flow levels. For slope-discharge erosion laws like the ones used in this study, the effect of flow variability can usually be subsumed into the erosion parameter K (Tucker, 2004). For erosion laws that include a threshold term, however, one might expect that fixed-discharge and variable-discharge laws would predict different behavior. The former implies the existence of locations where the threshold is never exceeded, whereas the latter allows erosion to occur at some rate everywhere simply because the threshold will always be exceeded for some fraction of time. Despite this, we find little difference between fixed-discharge and stochastic-discharge models even when a threshold is included (Figure 8, red hexagon). The similarity in behavior presumably reflects our use of a smoothed (as opposed to fixed) threshold function, which also allows some degree of erosion to occur at all locations. The similarity in performance between stochastic and deterministic model variants implies that one can include the effects of flow variability by incorporating them into a single lumped parameter, without needing the extra parameters and degrees of freedom that the stochastic models require.

5.1.4. Model Structure Improvements

Based on our calibration results, there are two primary artifacts that present targets for improving models. First, many models exhibit long, linear incision features in the southern part of the model domain where the bedrock-dominated upland area meets the flat-lying till plateau (see, for example, Figure 7c). All models use the common D8 method to direct and accumulate surface water. This approach to surface hydrology does not handle divergent areas accurately, and is the likely culprit behind the observed incision features. These features are thus artifacts of an oversimplified representation of surface hydrology. This finding is broadly consistent with a study by Hancock et al. (2010), who compared SIBERIA (a numerical model with single-direction flow routing) with CAESAR (a cellular model with multiple flow direction) in the context of Tin Camp Creek catchment in Australia. They found that some areas of the catchment had enhanced incision in SIBERIA model runs, possibly due to overprediction of flow convergence as a result of the D8 routing algorithm. We recommend that future work consider alternative surface water routing schemes as part of the model structure space.

A common challenge in comparing numerical models is that the programs often differ in multiple ways, which makes it difficult to isolate individual processes or effects. One potential benefit of developing models within a consistent framework such as the Landlab Toolkit is that models can be constructed in such a way that only one element changes at a time (e.g., use of standard D8/Steepest Descent flow routing or using the OverlandFlow shallow water flow component of J. M. Adams et al., 2017). This permits, for example, the isolation of the flow routing effects so as to shed light on when different types of flow routing change model analysis results (e.g., Shelef & Hilley, 2013).

We note that none of the models that explicitly treat soil (Sa) perform especially well. Examination of the resulting modeled topographic residual (for example, for model BasicRtSa (C00), Figure 9) indicates that these models under-predict erosion and widening of the major valley side slopes. Model output indicates that adjacent to these channels is a thin soil layer. Recall that in the Sa model variants, downslope material transport is limited by the thickness of the available soil layer, and the rate of production of this soil layer is itself limited. We interpret the relatively poor performance of models with a dynamic soil layer to indicate that even with the highest justifiable soil production rates, soil cannot form fast enough in this model to keep up with rapid stream incision. The model fails to account for the fact that the glacial material is capable of failing and moving downslope without first being weathered into soil. In retrospect it may seem obvious that it is necessary to permit till to move in this environment—however, it is only through examining the calibration results that we were able to identify that no reasonable soil production rate was sufficient in this context. A clear next step to improve this model would be to allow soft lithologies such as till to move

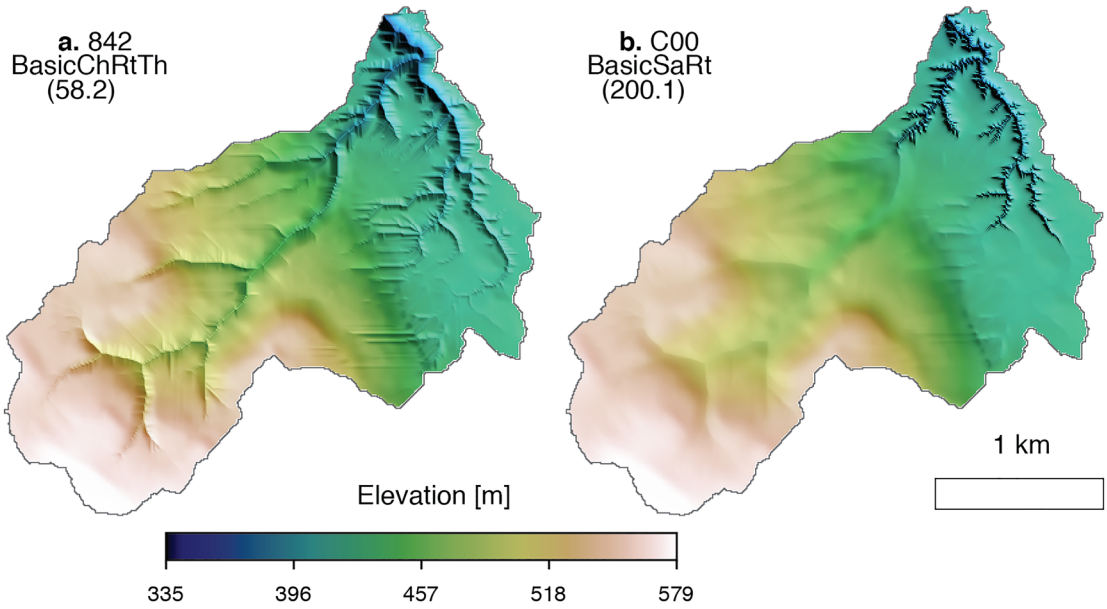

Figure 9. Comparison of end-of-model run topographic elevation for (b) poorly performing model C00 BasicRtSa with (a) the best performing model 842 BasicChRtTh. In BasicRtSa model hillslopes do not respond to the deeply incised main channel because mobile regolith is not produced fast enough. Little incision occurs in the upper portions of the watershed. The objective function value is given in parentheses.

downslope through hillslope processes, and configure the model with two transport coefficients: a larger one that applies to soil, and a smaller one that permits the till to move without being first converted to soil.

The failure of Sa models is an example of the general utility of calibration in identifying weaknesses in our knowledge of geomorphic process or model representation. In this case, our process representation of an explicit soil layer was well intentioned but flawed. Yet it was only through calibrating this model that we identified this flaw.

5.2. Assessment of Model Ranking Based on Validation

A validation effort serves as an independent check on the calibration. If all models performed equally well in calibration and in validation, they would plot on the 1:1 line in Figure 10. Models do not plot on this line, but are offset and almost parallel to it. We can draw two primary conclusions from these results. First, the validation results support the relative rank-ordering of the models identified by calibration. Second, the clear distinction in calibration score between BasicChRtTh (842) and the remaining Rt models is not as clear in the validation scores.

We made validation model runs for all 34 models on 5 initial condition topographies described by Barnhart et al. (2020b, their section S3). Examination of the validation results by initial condition (Figure S36) indicates that all models performed much worse on the postglacial initial topography with 0% etching. This makes sense when we compare the topography in the two drainage basins—the validation basin has a large, very flat region in the lower portion of the watershed. Small changes in the placement of the main channel in this area therefore result in large objective function values. For this reason we excluded the 0% etching validation results from further analysis.

The best performing models are those that distinguish between rock and till. However, in validation, the clear benefit of BasicChRtTh over the remaining Rt models does not persist. Model BasicDdRt (808) performs best overall in validation, followed by BasicRtTh (802) and BasicChRtTh (842). The fact that each of these models includes an erosion threshold supports the conclusion that a threshold is an important element in a successful model for this domain. The validation results, however, do not provide support for the conclusion that model BasicChRtTh is significantly better than the other 800-variant models.

A plot of modern versus modeled topographic residual for each of the validation model runs assists with interpretation of the offset between the 1:1 line and the validation results in Figure 10 (Figure 11). Figure 11 can be contrasted with Figure 5. Almost all validation model runs show a similar pattern of misfit between

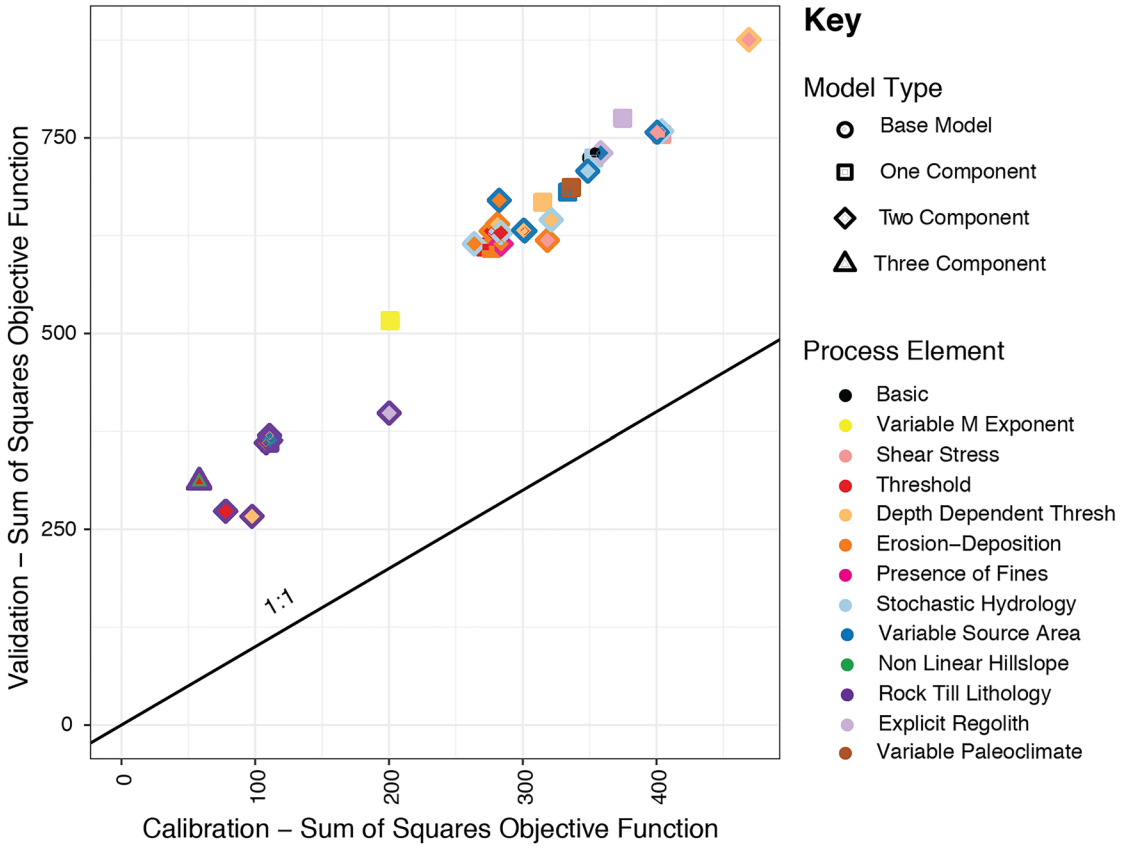

Figure 10. Comparison of model performance in the calibration and validation watersheds. Application of the calibrated models to the validation domain does not substantially change the model ranking.

modern and modeled topography in the upper elevations of the watershed. These model runs produce topography in the upper portions of the watershed that is closer to the modern validation domain than the respective domain in the calibration watershed. In contrast, there is not sufficient incision in the lower portion of the validation watershed, and the misfit is greater than that in the calibration watershed.

Examination of the calibration and validation results from model BasicChRtTh reveals the main reasons the model performed poorly in the validation domain (Figure 12). The topographic residual indicates reasonably good model fit in the area of the main channel that drains the bedrock-dominated uppermost portion of the watershed (which runs from the outlet to w to x in Figure 12). The next largest channel (at the point marked y) has incised deeply enough, but the river valley has not widened sufficiently, and the remainder of the channels (around the point marked z) have not incised sufficiently and do not have wide enough valleys. We speculate that this is because many of the channels in the area marked z have very low drainage area and are not able to incise sufficiently in the 13 ka duration. As valley widening at this site occurs primarily due to hillslope response to channel incision, the V-shaped valleys that cut the till plateau are not sufficiently wide. It is tempting to assume that the discrepancy arises because the modeled valley bottom width is narrower than the actual valley bottom width. However, the observed-valley bottom width is on the order of 10 m or less (similar to the model grid resolution).

An additional consideration, motivated by the observation that the validation watershed has much greater drainage density than the calibration watershed in the till plateau areas, is that runoff generation mechanisms are different in the two areas. This may reflect the presence of an alluvial fan deposit in the calibration watershed (LaFleur, 1979), which may be more permeable and generate less runoff than the glacial sediments.

While our objective function was designed to be equivalent across watersheds, and the validation watershed was chosen as the most similar watershed to Upper Franks Creek, the offset between the calibration and validation performance is an indication that improvements to our objective function are needed. Small

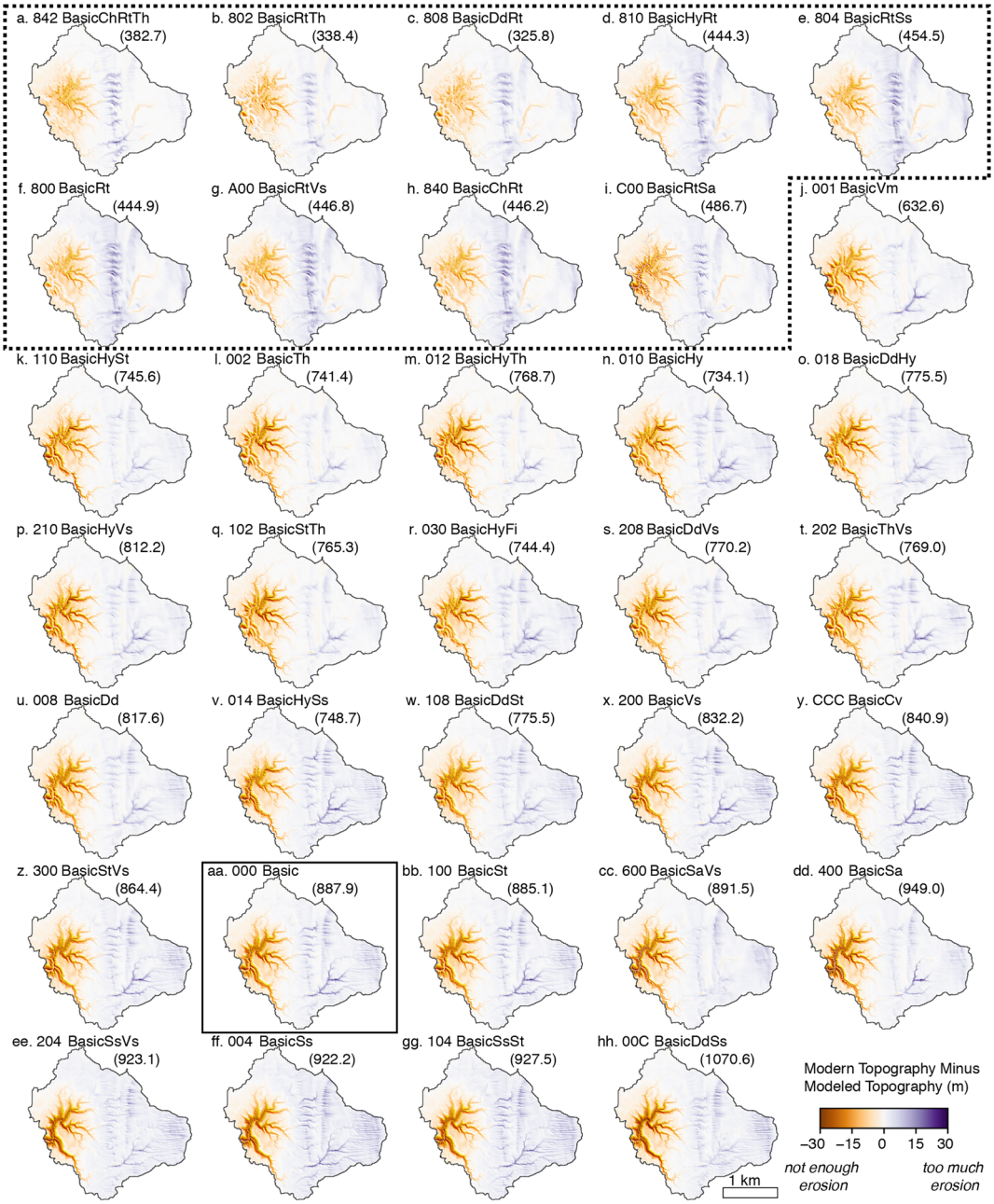

Figure 11. Maps of model fit (observed minus simulated) for all 34 successfully calibrated models run on the validation domain with the 7% etching initial condition. Models are ordered based on calibration ranking from best (a) to worse (hh), and the value in parentheses provides the validation objective function value. The dashed line outlines models with the Rt element and the solid line identifies model Basic.

differences, such as a larger portion of flat-lying areas or slightly different runoff generation characteristics, may always present a challenge in landscape evolution model validation.

5.3. Lessons From the Current Objective Function

The "elevation patch" objective function defined in section 3.3 was developed to be portable to multiple watersheds, take advantage of the beneficial mathematical properties of the $L2$ norm and least squares calibration, and produce sensible relative model rankings. It was for similar reasons that Skinner et al. (2018) used a stream-order-based spatial aggregation. While the elevation patch objective function proved sufficient for this initial application, using it across such a large model space exposed areas for future improvement.

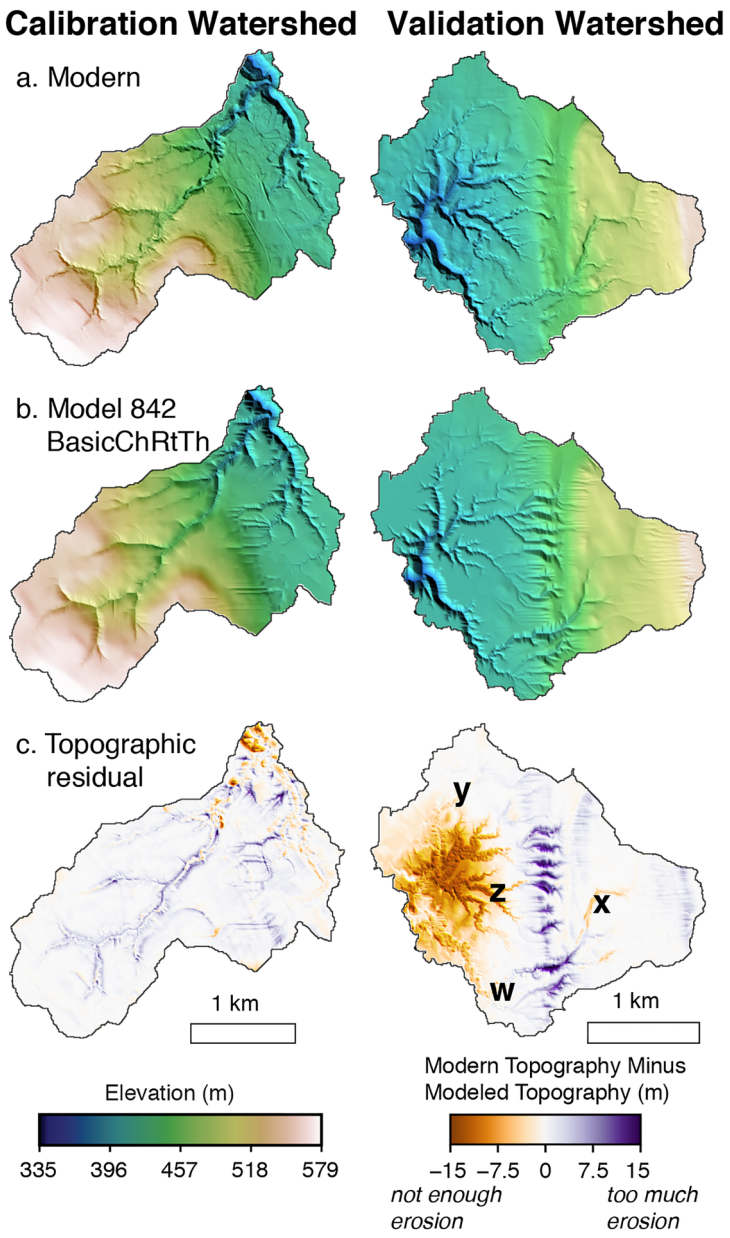

Figure 12. Comparison of calibration (left column) and validation(right column) watersheds for model 842 BasicChRtTh. Modern topography (a), end-of-model run topography (b), and topographic residual (c) illustrate the spatial patterns of model performance. Letters w, x, y, and z refer to locations discussed in the text.

After discussing the results of an initial effort to construct an alternative objective function based on statistical metrics of watershed topography, we enumerate lessons learned from application of the elevation patch objective function.

5.3.1. Objective Function Based on Topographic Metrics

Exploratory calibrations using the Gauss-Newton method and a preliminary objective function composed of topographic metrics such as the mean elevation and hypsometric integral (see Barnhart et al., 2020b, their section 6 for full list) failed to meet Criterion 2. This criterion requires that model ranking based on the objective function not be in conflict with model ranking based on expert assessment of simulated equivalents. We discuss this failure in the spirit of supporting future development of objective function components. Poor performance of an objective function composed of topographic metrics may arise, in part, from the weighting of individual metrics. In order to use the metrics in the context of formal calibration, it is necessary to

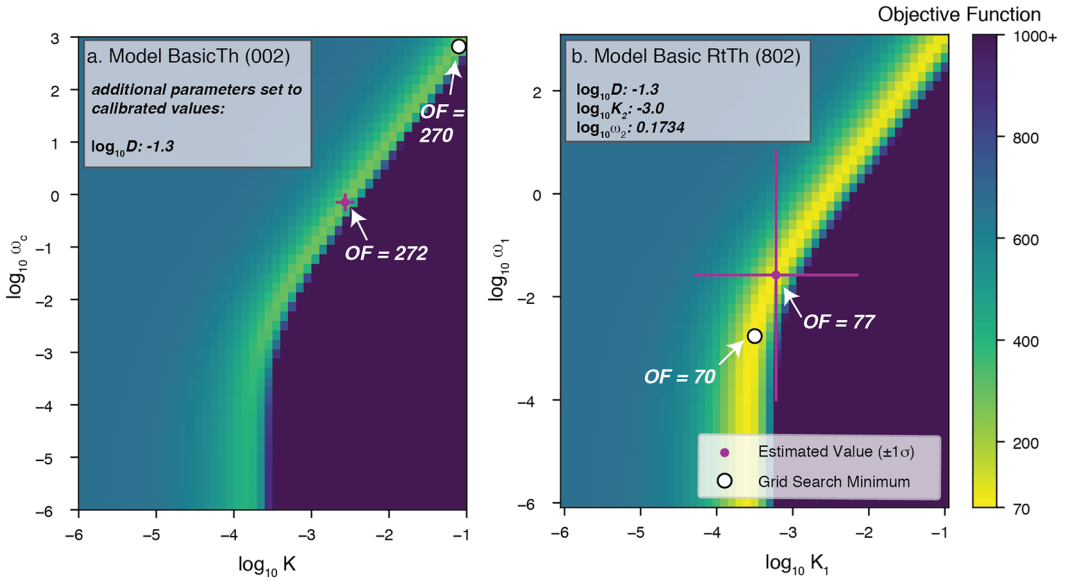

Figure 13. Slices through the objective function surface of models 002 BasicTh (a) and 802 BasicRtTh (b) demonstrating correlation between erodibility parameters (x axis) and erosion thresholds (y axis). The slices are taken at the minimum values for the other model parameters.

define weights to combine them into a single objective function. Each weight reflects the sources of error in the observation, including measurement error and model error (Hill & Tiedeman, 2007, their Guideline 6, p. 291). Constructing the objective function also requires bringing multiple observations with different units and/or dimensions onto a unified scale. In our development of this preliminary objective function we struggled to identify appropriate weights to combine the topographic metrics into an objective function. Multiobjective optimization approaches may provide a promising approach to identifying suitable weights. They, however, are too computationally intensive for our analysis of 37 landscape evolution models.

5.3.2. Objective Function Based on Topographic Difference

Next we consider the objective function based on direct topographic difference. First, it has properties that make application of gradient-based optimization such as nonlinear least squares challenging. For example, the objective function has many small local minima (Figures 3 and 4). In use cases like ours, in which it is not possible to prove convexity for any of our models, the possibility of local minima will persist. It is beyond the scope of this work to fully diagnose the origin of these minima. These local minima may be a property of many landscape evolution models, they may be a property of the components of the objective function (P_j in equation (4)), they may be related to certain numerical approximations (such as the route-to-one flow accumulation algorithm), or they may result from mathematical artifacts (Clark & Kavetski, 2010; Kavetski & Kuczera, 2007; Kavetski & Clark, 2010).

Second, the models have a number of parameters that are process-critical but have little influence on the objective function. An example of such a parameter is the linear diffusion coefficient D. Examination of parameter estimates from calibration indicates that this parameter was associated with large uncertainties when using a least squares method for calculating confidence intervals. Figures 3 and 4 indicate that the bottom of the objective function surface is relatively flat. Future work should identify objective function components that are more sensitive to D, as it is almost always the only model parameter that controls hillslope evolution.

5.3.3. Impact of Parameter Correlation

A third issue pertains to correlation between process parameters. Investigation of the objective functions of models 002 BasicTh and 802 BasicRtTh, for example, indicate that some calibration parameters are highly correlated with one another (Figure 13). Similar patterns of parameter correlation in objective function surfaces have been presented in prior studies in geomorphology (Croissant & Braun, 2014; Pelletier et al., 2011; Tomkin et al., 2003).

Both model BasicTh and BasicRtTh use erosion thresholds, such that fluvial incision is given as (for model 002)

$$E = KA^{1/2}S - \omega_c \left(1 - e^{\frac{KA^{1/2}S}{\omega_c}}\right) . \quad (10)$$

Examination of equation (10) reveals that one should expect correlation between K and ω_c: a change in either of these parameters could produce the same change in E. We formulated the equation in this way based on existing theory and because there is existing literature that constrains the values of K and ω_c (and related parameters; see Barnhart et al., 2020c). However, when identifying parameters for a calibration algorithm to estimate, it is helpful to minimize parameter correlation because when two parameters are highly correlated the calibration algorithm can get the same objective function value by changing the two parameters in concert. In effect, this parameter inter-correlation means that only one parameter can be estimated. One solution is to use a dummy parameter. For example, one could define a dummy factor F_K such that $K = F_K\omega_c$. This dummy variable effectively represents the portion of K that is independent of ω_c, such that both F_K and ω_c can be estimated through calibration. We recommend that future calibration efforts assess the inter-related nature of model parameters and seek to reduce unnecessary correlation.

The impact of covariance is further demonstrated by comparing the calibrated parameter values for models 800 and 802. Model 800 includes an erodibility for till K_1 and rock K_2. Calibration of this model yields values of $\log_{10}K_1 = -3.6$ and $\log_{10}K_2 = -7.1$ (Table S28), indicating that rock is much more difficult to erode than till. Model 802 adds erosion thresholds for till ω_1 and for rock ω_2. The calibration results for this model indicate that the erodibility values are more similar ($\log_{10}K_1 = -3.2$ and $\log_{10}K_2 = -3$) while the difference between the two substrates is taken up by the thresholds ($\log_{10}\omega_1 = -1.57$ and $\log_{10}\omega_2 = -0.17$, Table S29).

5.3.4. Assessing Nonlinear Least Squares Assumptions

Finally, we note that the objective function does not conform to three assumptions underlying nonlinear least squares calibration methods. For reasons described below, we do not think that this is problematic for our conclusions. However, we describe the flaws in our objective function to enable future improvements to similar efforts.

First, the shape of the objective function in the vicinity of the global minimum for D in Figure 4 is rather flat. This results in large parameter confidence intervals because these intervals are based on the numerical estimate of the Hessian (second derivative matrix of the objective function with respect to the parameters). A flat objective function surface is not well approximated by the first-order Taylor series expansion used to numerically calculate the Hessian in nonlinear least squares methods. Second, the shape of the objective function is not symmetric, particularly in K. This implies that linear confidence intervals may be an insufficient approximation of parameter uncertainty. Finally, the shape of the minimum (yellow areas in Figures 3 and 13) implies that estimated parameters are not always linearly separable.

Despite the above issues, the nonlinear least squares is a well established and well-vetted statistical method that is commonly used to draw inferences even when all underlying assumptions are not met. The result that all rock-till models outperform any other model further supports our geomorphic inference. Based on our interrogation of the objective function properties, we recommend that results dependent on calibrated parameter estimates are assessed to determine whether assumptions are valid. When comparing observed and modeled landforms, non-Gaussian objective functions are likely to be the norm rather than the exception. Successful calibration therefore relies on methods that can handle such complex surfaces.

Formal model analysis reveals strengths and weaknesses in model construction and assumptions, as well as the elements of the objective function. Based on the results presented in studies such as Clark and Kavetski (2010), we anticipate that resolving these issues lies in the domain of Earth surface processes rather than in statistics or inverse theory. We presently have no basis to argue that the sum of squared residual form of the objective function is flawed for these applications. Future efforts using model analysis methods should work toward identifying objective function elements and models that result in smoother and more robust objective functions. For example, Furbish (2003) explored a potential improvement and advocated for a promising approach: comparing models and observations using two coupled state variables together, such as soil depth and elevation.

5.4. Implications for Other Locations and Models

The extent to which our results will translate to other catchments, spatial scales, and temporal durations is not known. While the details of which aspects of model complexity improve model performance many not be consistent from one location or scale to another, there are two major aspects of our results that are likely to transfer to other applications in Earth surface process modeling.

First, the multimodel approach provides a formal avenue for testing hypotheses related to which geomorphic transport formulae and/or other aspects of model development best represent the dynamics of long-term landscape evolution. We are not the first to use such an approach (e.g., Doane et al., 2018; Hancock et al., 2010; van der Beek & Bishhop, 2003). A major benefit of developing our application within the Landlab Toolkit framework is that it is relatively easy to extend our model set to include other modifications to governing equations. While the set of models we consider does not cover all possible choices in process representation or all possible geomorphic transport formulas that have been considered in landscape evolution modeling (see Tucker & Hancock, 2010, for a review), our choices cover many of the most commonly used principles.

Our application of a hybrid surrogate-based global and complex model gradient-based optimization method is a promising portable approach for applications with long simulation times and objective functions with unknown properties (e.g., smoothness, local minima). However, further work on refining objective functions and optimization techniques in Earth surface processes modeling is needed to determine which objective functions and optimization methods are consistently successful.

Second, we expect that our finding of nonlinearities in model structure and parameter space will be the rule and not the exception, especially in the case of simulation timescales that are shorter than typical process adjustment timescales. The results of Barnhart et al. (2020b) indicating highly nonlinear effects in sensitivity analysis are consistent with existing sensitivity analyses in Earth surface process modeling (e.g., Shobe et al., 2018; Skinner et al., 2018; Temme & Vanwalleghem, 2016; Ziliani et al., 2013). The challenges of parameter covariance described by Figure 13 are similar to objective functions shown by Tomkin et al. (2003) and Croissant and Braun (2014). This is likely the result of the presence of thresholds in many geomorphic transport formulae.

6. Conclusions

We present a novel study in the application of the methods of model analysis to landscape evolution modeling. Using a hierarchical suite of alternative models designed to efficiently explore a high-dimensional model structure space we identified the model structure permutations that improve simulation performance.

Calibration of multiple alternative models using the same objective function and formal calibration algorithms discriminates meaningfully between alternative models, and thereby reveals which geomorphic process permutations add value when simulating a system. Validation is needed to provide an independent check; our validation results show that the benefits of the rock-till differentiation and erosion threshold permutations persist when the calibrated models are applied to a new watershed. The ranking of other process elements is less consistent.

Application of multimodel analysis in landscape evolution is both a useful tool and a philosophical approach. Given the state of uncertainty about the form of governing equations and appropriate ways to simplify them for use on geologic timescales, multimodel analysis provides a framework for true hypothesis testing of landscape evolution theory.

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 supported by the University of Colorado Boulder "PetaLibrary." Discussion with Christopher Miller substantially improved this manuscript. 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., 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., 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).

Adams, J. M., Gasparini, N. M., Hobley, D. E. J., Tucker, G. E., Hutton, E. W. H., Nudurupati, S. S., & Istanbulluoglu, E. (2017). The Landlab v1. 0 overlandflow component: A python tool for computing shallow-water flow across watersheds. *Geoscientific Model Development*, *10*(4), 1645–1663. https://doi.org/10.5194/gmd-10-1645-2017

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akadémiai Kiado.

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

Andrews, D. J., & Bucknam, R. C. (1987). Fitting degradation of shoreline scarps by a nonlinear diffusion model. *Journal of Geophysical Research*, *92*(B12), 12,857–12,867. 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 multi-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*(42), 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: 1. Conceptualization and sensitivity analysis. *Journal of Geophysical Research: Earth Surface*, *125*, e2018JF004961. https://doi.org/10.1029/2018JF004961

Barnhart, K. R., Tucker, G. E., Doty, S., Shobe, C. M., Glade, R. C., Rossi, M. W., & Hill, M. C. (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

Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. *Journal of Hydrology*, *105*(1-2), 157–172. https://doi.org/10.1080/02626667909491834

Beven, K. (2002). Towards a coherent philosophy for modelling the environment. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *458*(2026), 2465–2484. Retrieved from http://www.jstor.org/stable/3067324

Bishop, P. (2007). Long-term landscape evolution: Linking tectonics and surface processes. *Earth Surface Processes and Landforms*, *32*, 329–365. https://doi.org/10.1002/esp.1493

Bras, R. L., Tucker, G. E., & Teles, V. (2003). Six myths about mathematical modeling in geomorphology. In P. Wilcock, & R. Iverson (Eds.), *Prediction in geomorphology* (pp. 63–79). Washington, DC, USA: American Geophysical Union. https://doi.org/10.1029/135GM06

Buehler, E. J., & Tesmer, I. H. (1963). Geology of Erie County, New York. *Buffalo Society of Natural Science*, *21*(3), 118.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, USA: Springer-Verlag. ISBN 0387953647.

Clark, M. P., & Kavetski, D. (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. *Water Resources Research*, *46*, W10510. https://doi.org/10.1029/2009WR008894

Codilean, A. T., Bishop, P., & Hoey, T. B. (2006). Surface process models and the links between tectonics and topography. *Progress in Physical Geography*, *30*, 307–333. https://doi.org/10.1191/0309133306pp480ra

Coulthard, T. J. (2001). Landscape evolution models: A software review. *Hydrological Processes*, *15*, 165–173. https://doi.org/10.1002/hyp.426

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

Davis, W. (1892). The convex profile of bad-land divides. *Science*, *20*(508), 245. https://doi.org/10.1126/science.ns-20.508.245

Dennis, J. E. Jr., Gay, D. M., & Walsh, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, *7*(3), 348–368. https://doi.org/10.1145/355958.355965

Dennis Jr, J. E., & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations* (Vol. 16). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611971200

Dietrich, WE, Bellugi, D. G., Sklar, L. S., Stock, J. D., Heimsath, A. M., & Roering, J. J. (2003). Geomorphic transport laws for predicting landscape form and dynamics. In P. Wilcock, & R. Iverson (Eds.), *Prediction in geomorphology* (pp. 103–132), Geophysical Monograph Series. Washington, DC: AGU. https://doi.org/10.1029/135GM09

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

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. W., Hill, M. C., & 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

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

Gilbert, G. (1877). *Report on the geology of the Henry Mountains* (Vol. 160). Washington, DC: U.S. Geographical and Geological Survey of the Rocky Mountain Region. https://doi.org/10.3133/70039916

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

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. London, UK: Academic Press.

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

Grayson, R. B., Moore, I. D., & McMahon, T. A. (1992). Physically based hydrologic modeling: 2. Is the concept realistic? *Water Resources Research*, *28*(10), 2659–2666. https://doi.org/10.1029/92WR01259

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., 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., & 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

Hanks, T. C. (2000). The age of scarplike landforms from diffusion-equation analysis. *Quaternary Geochronology: Methods and Applications*, *4*, 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

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., & 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

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*(4), 455–492. https://doi.org/10.1023/A:1008306431147

Kavetski, D., & Clark, M. P. (2010). Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research*, *46*, W10511. https://doi.org/10.1029/2009WR008896

Kavetski, D., & Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resources Research*, *43*, W03411. https://doi.org/10.1029/2006WR005195

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362

Klemeš, V. (1986). Dilettantism in hydrology: Transition or destiny? *Water Resources Research*, *22*(9S), 177S–188S. https://doi.org/10.1029/WR022i09Sp0177S

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

Lane, S. N., & Richards, K. S. (2001). The 'validation' of hydrodynamic models: Some critical perspectives. *Model validation: Perspectives in hydrological science*, *413*, 439.

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

Martin, Y., & Church, M. (2004). Numerical modelling of landscape evolution: Geomorphological perspectives. *Progress in Physical Geography*, *28*, 317–339. https://doi.org/10.1191/0309133304pp412ra

McKenna, S. A., & Poeter, E. P. (1995). Field example of data fusion in site characterization. *Water Resources Research*, *31*(12), 3229–3240. https://doi.org/10.1029/95WR02573

Molnar, P. (2001). Climate change, flooding in arid environments, and erosion rates. *Geology*, *29*(12), 1071–1074. https://doi.org/10.1130/0091-7613(2001)029<1071:CCFIAE>2.0.CO;2

Pazzaglia, F. J. (2003). Landscape evolution models. *Developments in Quaternary Sciences*, *1*, 247–274. https://doi.org/10.1016/S1571-0866(03)01012-1

Pelletier, J. D. (2013). Fundamental principles and techniques of landscape evolution modeling. In J. F. Shroder (Ed.), *Treatise on geomorphology* (pp. 29–43). San Diego, CA, USA: Elsevier Inc. https://doi.org/10.1016/B978-0-12-374739-6.00025-7

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, JD, 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*, 191–24. 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., & 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.3302

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*, 367. 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

Poeter, E. P., & McKenna, S. A. (1995). Reducing uncertainty associated with ground-water flow and transport predictions. *Groundwater*, *33*(6), 899–904. https://doi.org/10.1111/j.1745-6584.1995.tb00034.x

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

Shelef, E., & Hilley, G. E. (2013). Impact of flow routing on catchment area calculations, slope estimates, and numerical simulations of landscape development. *Journal of Geophysical Research: Earth Surface*, *118*, 2105–2123. https://doi.org/10.1002/jgrf.20127

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

Skinner, C. J., Coulthard, T. J., Schwanghart, W., Van De Wiel, M. J., & Hancock, G. (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

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics-Theory and Methods*, *7*(1), 13–26. https://doi.org/10.1080/03610927808827599

Tarantola, A. (1987). *Inverse problem theory: Methods for data fitting and model parameter estimation* (p. 644). New York, NY, USA: Elsevier Science Pub. Co. ISBN 0444427651.

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., Schoorl, J. M., Claessens, L., & Veldkamp, A. (2013). Quantitative modeling of landscape evolution. In J. F. Shroder (Ed.), *Treatise on geomorphology* (pp. 180–200). San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-374739-6.00039-7

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*, 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

Valters, D. (2016). Modelling geomorphic systems: Landscape evolution. *Geomorphological Techniques*, *12*, 1–24. https://doi.org/10.13140/RG.2.1.1970.9047

van der Beek, P., & Bishhop, 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*, 2309. https://doi.org/10.1029/2002JB002125

Willgoose, G. R. (2005). Mathematical modeling of whole landscape evolution. *Annual Review of Earth and Planetary Sciences*, *33*, 14.1–14.17. https://doi.org/10.1146/annurev.earth.33.092203.122610

Willgoose, G. R., Bras, R. L., & Rodriguez-Iturbe, I. (1991). 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., & Hancock, G. R. (2011). Applications of long-term erosion and landscape evolution models, *Handbook of erosion modelling* (chap. 18, pp. 339–359). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444328455.ch18

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, RA (2018). *Phase 1 erosion studies: Study 1–Terrain analysis* (Tech. Rep.). West Valley Erosion Working Group. Retrieved from https://wvphaseonestudies.emcbc.doe.gov/Documents/EWG%20Final%20Study%201%20Report_Vol%201_2.21.18.pdf

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