Oliver Korup

Bayesian geomorphology

Suggested citation referring to the original publication:
Earth surface processes and landforms : the journal of the British Geomorphological Research Group 46 (2020) 1, pp. 151 - 172
DOI: https://doi.org/10.1002/esp.4995
ISSN: 0197-9337, 1096-9837

Journal article | Version of record

Secondary publication archived on the Publication Server of the University of Potsdam:
Zweitveröffentlichungen der Universität Potsdam : Mathematisch-Naturwissenschaftliche Reihe 1348
ISSN: 1866-8372
URN: https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-539892
DOI: https://doi.org/10.25932/publishup-53989

EARTH SURFACE PROCESSES AND LANDFORMS
Earth Surf. Process. Landforms **46**, 151–172 (2021)

Published online 21 September 2020 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/esp.4995

State of Science

Bayesian geomorphology

Oliver Korup*
Institute of Environmental Science and Geography, Institute of Geosciences, University of Potsdam, Potsdam D-14476, Germany

Received 5 May 2020; Revised 26 August 2020; Accepted 27 August 2020

*Correspondence to: Oliver Korup, Institute of Environmental Science and Geography, University of Potsdam, D-14476 Potsdam, Germany, Email: korup@uni-potsdam.de

ESPL
Earth Surface Processes and Landforms

SUMMARY: The rapidly growing amount and diversity of data are confronting us more than ever with the need to make informed predictions under uncertainty. The adverse impacts of climate change and natural hazards also motivate our search for reliable predictions. The range of statistical techniques that geomorphologists use to tackle this challenge has been growing, but rarely involves Bayesian methods. Instead, many geomorphic models rely on estimated averages that largely miss out on the variability of form and process. Yet seemingly fixed estimates of channel heads, sediment rating curves or glacier equilibrium lines, for example, are all prone to uncertainties. Neighbouring scientific disciplines such as physics, hydrology or ecology have readily embraced Bayesian methods to fully capture and better explain such uncertainties, as the necessary computational tools have advanced greatly. The aim of this article is to introduce the Bayesian toolkit to scientists concerned with Earth surface processes and landforms, and to show how geomorphic models might benefit from probabilistic concepts. I briefly review the use of Bayesian reasoning in geomorphology, and outline the corresponding variants of regression and classification in several worked examples. © 2020 The Authors. Earth Surface Processes and Landforms published by John Wiley & Sons Ltd

KEYWORDS: Bayes' rule; probability; uncertainty; prediction

Acknowledging Uncertainty in Geomorphology

Geomorphology, like any other science, is much about learning. We learn by collecting data and thus reducing, step by step, our uncertainties about the shapes and dynamics of landscapes that surround us. Bayesian theory offers the tools to measure these uncertainties, and this article intends to introduce some of the underlying concepts. Fields such as ecology, hydrology, meteorology or seismology have been embracing Bayesian methods for some time now (Dose and Menzel, 2004; Seidou *et al*. 2006; Silva *et al*. 2015; Jo *et al*. 2016), although geomorphologists have invested somewhat less in this approach. Yet rarely do we encounter data, problems or predictions that are free of uncertainty. Often enough, we phrase our uncertainty in questions. 'How do bedform patterns arise?', 'Time to abandon the Manning equation?' or 'Long-term river meandering as a part of chaotic dynamics?' are among nearly 60 other questions that serve as titles of articles published in *Earth Surface Processes and Landforms* in the past 10 years. Any rhetorical intentions aside, the authors of these studies may have wished to communicate some struggle with their data or their possible interpretation. Anyone who has mapped landforms in the field or from air photos will acknowledge that some ambiguity is always part of the process; the same goes for inferring how the thickness of soil or sediment varies between a handful of pits, trenches or outcrops. The field of geostatistics has grown largely from this necessity of interpolating between point measurements in landscapes, and modern interpolation models are designed to predict both in space and time (Cressie and Wikle, 2011). The theory of *fuzzy logic* admits that the quantity we wish to learn may be partly imprecise (Zadeh, 2006), and provides a mathematical context for visualizing uncertainty on maps (Wheaton *et al*. 2010).

In the early 21st century, data are ubiquitous and information has become a powerful currency. The field of geomorphology is experiencing a steep rise in the amount and diversity of freely available data, let alone the many new technologies to use them. To take advantage of this development, we need the tools and skills to handle, understand and interpret these data. A host of new trends such as *big data*, *machine learning*, *data science* and *Bayesian reasoning* have adopted concepts from both statistical and computational science, and now offer myriad algorithms and solutions to the problem of using data to predict under uncertainty (Witten *et al*. 2011; Barber, 2012). On the flip side, this uncertainty also means dealing with sparse, inhomogeneous or inaccurate data, a situation that geomorphologists often face. Bayesian reasoning offers a formal and consistent way to learn from data – regardless of whether information is sparse or abundant – while measuring explicitly what we have learned compared to our previous knowledge. Bayesian data analysis uses probability distributions to fully document and gauge this learning process. This approach sets Bayesian

methods apart from machine learning algorithms such as gradient boosting, random forests or artificial neural networks that can perform extremely well, though at the cost of remaining black boxes with little insight into how they arrived at their results (Breiman, 2001; LeCun *et al*. 2015). Arguably, Bayes' rule might be one of the most influential scientific theorems of the 21st century (Efron, 2013), given that computational power is now sufficient to apply probabilistic reasoning to real-world problems with large datasets and many parameters. This development is important to us, because as geomorphologists we wish to make predictions, be it whether or how a landform develops, or when or how the rate of a process changes.

In looking back, a large body of work deals with predicting geomorphic processes at all levels from theoretical concept to applied physics (Dottori *et al*. 2013; George and Iverson, 2014; Beven, 2015). Here *prediction* means making informed and objective statements about unobserved quantities. By this definition, prediction can refer to present or past events. *Unobserved* means that we lack data on the desired quantity. We can thus predict past or unrecorded events or those that happen right now, though beyond our sensory or instrumental reach. Many case studies have tested how varying data quality, resolution and the history of past landscape changes can affect predictions in geomorphology (Claessens *et al*. 2005; Stefanescu *et al*. 2012; Shikakura, 2014). One common strategy is to check how varying input values modulate the outputs of physically motivated or empirically calibrated models. This strategy is essential where predictions need to inform mitigation options against natural hazards (Ferreira *et al*. 2014; Beven *et al*. 2018a). In this context, bootstrapping and Monte Carlo simulations are popular methods to check whether model predictions are robust, and to propagate explicitly errors where we lack simple or closed analytical solutions (Rustomji and Wilkinson, 2008; Strenk and Wartman, 2011; Schwanghart *et al*. 2016). Hydrological research, for example, has been influenced heavily by Monte Carlo methods that aim at identifying models accepted as behavioural (Beven and Binley, 2014). Most studies of landslide susceptibility also make heavy use of probabilistic simulation (Guzzetti *et al*. 2005) in what is a classification problem: does a given set of terrain properties qualify a patch of hillslope as 'failure-prone' or 'stable', assuming that this distinction is mutually exclusive. Most of the classifiers used in these studies are based on probabilistic concepts; some have Bayesian roots (Mondini *et al*. 2013; Budimir *et al*. 2015). For example, *Naive Bayes* has become a popular alternative to logistic regression (Heiser *et al*. 2015; Kern *et al*. 2017). The term 'naive' alludes to a simplified use of Bayes' rule for what is an effective method for classifying data. Jensen *et al*. (2006) showed some simple applications of Bayes' rule for detecting rare geological events or quantities, especially for cases where the observations involve errors. *Bayesian weights-of-evidence* and variants thereof also feature in many studies of landslide susceptibility (Regmi *et al*. 2010; Berti *et al*. 2012), and use reasoning that is conceptually grounded in probabilistic inference. Many diagnostic statistics are based on Bayes' rule without necessarily being full Bayesian models. Charbonnier *et al*. (2018) showed one such example of how to use Bayesian reasoning for validating numerically simulated lahar runout with field measurements.

Still, many predictions that we use in geomorphology boil down to averages that we estimate from regression or classification models. While convenient, this reliance on mean estimates neglects variability of form and process. Probability distributions can conveniently encode this variability beyond a limited size of samples. Many phenomena like avalanche areas, flood discharge or the run-up heights of tsunamis appear to follow heavy-tailed distributions (Burroughs and Tebbens, 2005; Molnar *et al*. 2006), whereas the storage times of floodplain sediments, for example, appear to be exponentially distributed (Bradley and Tucker, 2013). Identifying suitable – and ideally physically relevant – distributions is key to forecasting the size, recurrence or patterns of form and process. This use of probability distribution paves the way for inferring, for example, the patterns of soil depths or permafrost (Boeckli *et al*. 2012). Probability distributions also form the pool from which to draw random samples that we input into mechanistic models. Herbert Einstein pioneered this approach in the 1940s by developing a probabilistic model of sediment transport in rivers (Dey, 2014). Turbulence in moving fluids remains a major challenge for predicting flow properties, and many mechanistic models resort to time- or depth-averaged approaches that carry some of the uncertainty by probabilistic measures of flow (Raffaele *et al*. 2018). Processes such as landsliding, sediment transport and soil perturbation can thus be treated probabilistically or tagged as *stochastic* (Benda and Dunne, 1997; Miller and Burnett, 2008; Bennett *et al*. 2014; Turowski and Hodge, 2017; Furbish *et al*. 2018). Still, these processes are bound to the laws of physics: it is our *uncertainty* – or incomplete knowledge – about these processes that motivates us to treat them as probabilistic.

In short, probability has been part of many geomorphic studies, measuring uncertainty in different nuances. This observation alone should motivate the use of a consistent and overarching framework for putting probability to good use in geomorphic prediction. In the following we explore how Bayesian reasoning can provide such a framework.

What Is Bayesian?

Why should we care about 'Bayesian' geomorphology? If anything, we should be curious about a way of thinking that has changed the way how scientists deal with data, models and interpretations (Efron, 2013). Bayesian reasoning is concerned with learning something new from data and existing knowledge. The way this reasoning works is that it combines both the known and the unknown by using probabilities. In this context, probability is a measure of uncertainty (Kadane, 2011). Broadly speaking, probability maps the (un)certainty about a given outcome to the unit interval, where 0 refers to the impossible, and 1 refers to the certain outcome. Regardless of this numerical constraint, two major, but different, interpretations of probability have fuelled a debate that may have made you reluctant to use Bayesian statistics (Gelman *et al*. 2004).

The conventional way of interpreting probability is that it represents the limit of an expected outcome based on an infinite number of repeat experiments. Much of classical statistical theory is built on this *frequentist* interpretation. In this theory, probability is an objective metric of a random variable that we have to study long enough to obtain better knowledge about its outcomes. If we had the chance (and endurance) to repeat scientific experiments endlessly, we would be able to approach with perfect precision (and accuracy) the probability of a given outcome. An alternative way is to interpret probability subjectively: it measures how uncertain you are about a quantity. Texts on Bayesian data analysis use 'believable' or 'credible' instead of the traditional 'statistically significant' to emphasize this view (Kruschke, 2015). Although we might try and minimize subjective elements in our research, they are part of learning and science. Fitting a line by eye to data points, discarding inverted radiocarbon ages as outliers, or preferring a power law over several other, equally plausible, models for sediment rating curves are just a few examples of subjective decisions in geomorphology.

The *Bayesian* approach requires us to express these subjective beliefs mathematically, and further recognizes that carrying out an infinite number of experiments is beside the point. In some cases, we may be limited to a single experiment. Consider doing fieldwork below an unstable rock cliff. What is the probability of being injured or even killed by a rock tumbling off the cliff face? The frequentist interpretation of probability might suggest you try out as many times as possible to find out. Clearly, it is undesirable to do this experiment more times than is necessary, if at all; a large number of trials is hard to achieve. Intuitively, we might think that it is unsafe to work beneath that cliff edge. Bayesian data analysis is geared to explicitly measure and account for this intuition (or any other relevant previous scientific knowledge) in a dedicated probabilistic term. Some interpretations go as far as to measure the degree of subjective belief with probability distributions. It is easy to see that this approach has met substantial critique given that we wish to do science objectively (Lavine, 2019). Yet Bayes' rule meets exactly that demand by requiring us to express our subjectivity in a mathematical, reproducible form. We can thus trace and reproduce any effects of our subjective beliefs on our results. Being wrong about an erosion rate, the age of a river terrace or the roughness of a channel bed is also part of geomorphic research. Probability distributions of our estimates can tell us how wrong we are likely to be.

Bayesian reasoning explicitly handles uncertainty and thus forms some of the essential building blocks of many algorithms and models currently used in machine learning or data science. The adjective 'Bayesian' acknowledges this root in the wider sense; a stricter view refers to methods that use Bayes' rule consistently throughout the entire process of data analysis. We start off with some basics of Bayes' rule. Consider a sandy beach on a tropical atoll littered by coral boulders from the nearby fringing reef. Without any rock cliffs, beach rock or uplifted fossil reefs anywhere in sight, we can safely assume that large waves pushed the boulders onto the beach. Assume that we know from earlier studies that 2% of boulders were moved onshore by tsunami waves. We can encode this *prior* knowledge as probability $P(T) = 0.02$. Now consider a new method M that uses the geometry and material properties of a boulder to reconstruct whether it is a tsunami deposit. This new method claims that it is 95% reliable. What is the probability $P(T|M)$ that a given boulder that you tested positively with this new method was dumped onto the beach by a tsunami? We read the '|' symbol as 'given' or 'conditional on', and answer this question by applying Bayes' rule. It is useful to keep in mind that this rule is firmly rooted in the axioms of probability theory and results directly from the definition of a conditional probability (Gelman *et al*. 2004):

$$P(T|M) = \frac{P(M|T)P(T)}{P(M)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{P(T,\, M)}{P(M)} \tag{1}$$

where $P(T|M)$ is the *posterior* probability of inferring the outcome of T ('was it a tsunami?') given that we already know the outcome of M ('a positive test result'). Thus the state of M is free of any randomness or uncertainty in $P(T|M)$ because we rely explicitly on knowing the result of M. Note that the posterior probability $P(T|M)$, or any other conditional probability, is completely neutral about which outcome happened first or whether one caused the other. The term *posterior* stands for what we learn about $P(T)$ after having updated our knowledge with respect to the known outcome of our test method M. The twist in Bayes' rule is that we can express the posterior $P(T|M)$ by its inversion $P(M|T)$, which is known as the *likelihood*. Think of the likelihood as a function that expresses how plausible you would expect an observation to be given a range of possible scenarios. Assuming that we knew that a tsunami moved our boulder, what are the chances that the method would detect that? If we were less sure about the tsunami origin, we should also admit a different origin. Accordingly, we call $P(T)$ the *prior*probability, and interpret this as our knowledge about the tsunami origin regardless of – or before seeing – test outcome M. The product of the likelihood and the prior is the *joint probability* of observing both T and M and denoted by $P(T,M)$. Finally, $P(M)$ is variably known as the *evidence*, *marginal likelihood*, or *average likelihood*. Whichever term you prefer, it remains a constant with respect to the posterior because this is conditional on the already known outcome of M ('a positive test result').

Having digested this theory, we can now use Bayes' rule to find that

$$P(T|M) = \frac{P(M|T)P(T)}{P(M)} = \frac{0.95 \times 0.02}{0.95 \times 0.02 + (1 - 0.95) \times (1 - 0.02)} = 0.279... \tag{2}$$

In plain words, the posterior probability with which we believe that a tsunami deposited this particular boulder, given that we tested the new method on it positively, is about 28%. This result can be surprising at first. The posterior probability is higher than the original 2%, because we obtained data by applying the new method to the boulder. However, the posterior probability is much lower than the 95% likelihood that the method is accurate, because tsunami boulders are rare on our beach. The posterior probability reconciles our prior knowledge and the likelihood by multiplication, and thus produces an updated compromise between the two. Having obtained a positive test result for a given boulder, we learn that our posterior belief about a tsunami source is now more than tenfold compared to our initial belief. We also see how prior knowledge adjusts or *penalizes* the likelihood by imposing weights in the form of a probability.

This simple example is instructive and perhaps counterintuitive at first, but it has limited scope (Figure 1). Bayes' rule offers more interesting and practical applications if we plug in probability distributions instead of discrete probabilities, and refer to the data $\mathscr{D}$ that we observe:

$$p(\kappa|\mathscr{D}) = \frac{p(\mathscr{D}|\kappa)p(\kappa)}{p(\mathscr{D})} \propto p(\mathscr{D}|\kappa)p(\kappa) \tag{3}$$

where κ is the quantity that we wish to learn, and $p(\cdot)$ identifies probability distributions (instead of discrete probabilities as in Equation (2)). Here κ stands for either

- the parameter(s) θ of a probability distribution or model that we assume has generated our data $\mathscr{D}$;
- unobserved outcomes $\hat{y}$ predicted by the model that we learned from $\mathscr{D}$;
- a model $\mathscr{M}$ that we believe to have generated the data $\mathscr{D}$; or

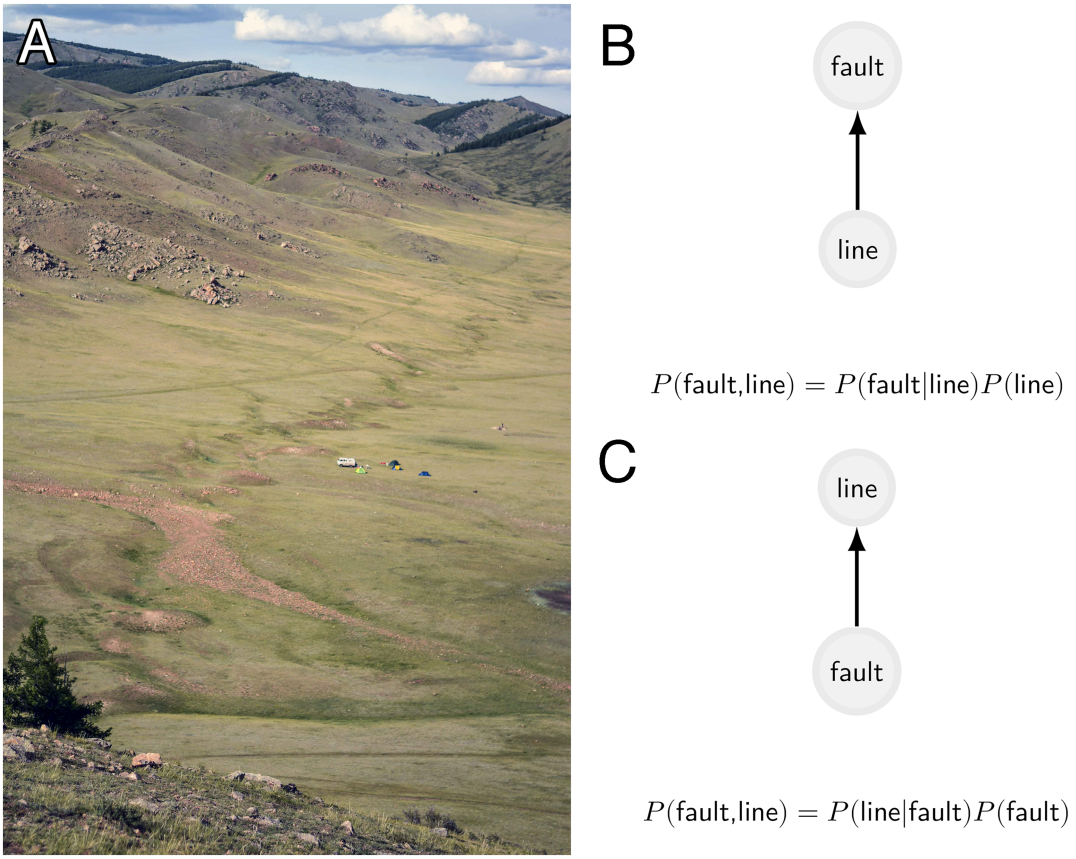

Figure 1. (A) Is that a fault line on this photo? A Bayesian geomorphologist might rephrase this question and wish to learn more specifically the probability of seeing a fault given the somewhat linear trace in the landscape, *P*(fault|line). (B) By definition, the joint probability of observing both fault and line *P*(fault, line) is the probability of observing a fault given the recognized line *P*(fault|line) times the probability of observing lines in any case, *P*(line); some lines in the photo can also be vehicle tracks. In this graph the bubbles are random variables, and arrows point from the conditioning to the conditional variable. (C) We can compute the same joint probability using *P*(line|fault) instead. Setting the right-hand sides of the equations in B and C equal, and solving for a given conditional probability gives Bayes' rule. Both *P*(line) and *P*(fault) may depend on your training as a geomorphologist or palaeoseismologist, and on how reliably you detect lineaments or faults independently of this particular setting (Bolnay Fault, Mongolia; vehicle and tents for scale). [Colour figure can be viewed at wileyonlinelibrary.com]

- a hypothesis $\mathscr{H}$ that we might find more credible than competing ones.

The posterior distribution $p(\kappa|\mathscr{D})$ shows what we learned about κ after having updated our previous knowledge $p(\kappa)$ with new data $\mathscr{D}$. The posterior is a probability distribution that is conditioned on the data, and hence also hinges on all assumptions and uncertainties that these data contain. In short, Bayes' rule links the unknown with the known in a convenient mathematical form that allows easily using past posteriors as new priors. Regardless of which quantity κ we wish to learn, the denominator in Equation (3), $p(\mathscr{D})$, is a constant that normalizes the product of likelihood and prior, and guarantees that the posterior is a proper probability distribution. In practice, we often compute the posterior from the product of likelihood and prior, and then *re-normalize* to obtain $p(\kappa|\mathscr{D})$. Re-normalizing means that we rescale the weights of all posterior outcomes such that they form a proper probability distribution.

Equation (3) states that we need to specify the joint distribution $p(\kappa, \mathscr{D})$, which we obtain from the product of the likelihood function(s) and the prior distribution(s). While the likelihood depends on both $\mathscr{D}$ and the assumed data-generating parameter(s), the prior expresses what we know already without any notion about $\mathscr{D}$. One useful application of this reasoning concerns the Bayesian calibration of radiocarbon ages by including prior knowledge about the stratigraphic context of the samples (Ramsey, 2009). The principles of stratigraphy tell us that organic samples taken from lower layers of a given section should be older than those from upper layers in general. Including this prior assumption into the calibration of many samples from a single section offers a consistent and logical way of eliminating spurious peaks in the multimodal probability distribution of calendar years. These peaks arise from local spikes of the calibration curve based on dated tree rings. We can thus achieve more consistent age models that take advantage of mapped stratigraphy and add value to detailed field observations (Blaauw *et al*. 2018).

To learn most effectively from the data, the prior should invite informed or expert knowledge (Nolde and Joe, 2013). We can gather initial information from experience or a synthesis of data in the literature. Packing this prior knowledge into a suitable probability distribution requires care if there are several, similarly adequate, options or if models consist of multiple levels (Gelman, 2006). Even experts may concede some uncertainty in their assessment or be prone to bias, so that eliciting prior knowledge and encoding it as a distribution systematically can be an iterative and demanding process (O'Hagan and Oakley, 2004; Kadane, 2011; Hassall *et al*. 2019). How to deal best with this process remains debated, though the process should recognize issues with few, missing, incomplete, inaccurate or non-stationary data (Beven *et al*. 2018b). In geomorphology we are often interested in learning about parameters that have physical limits, and we mostly have an idea about the shape of the distribution between these limits when

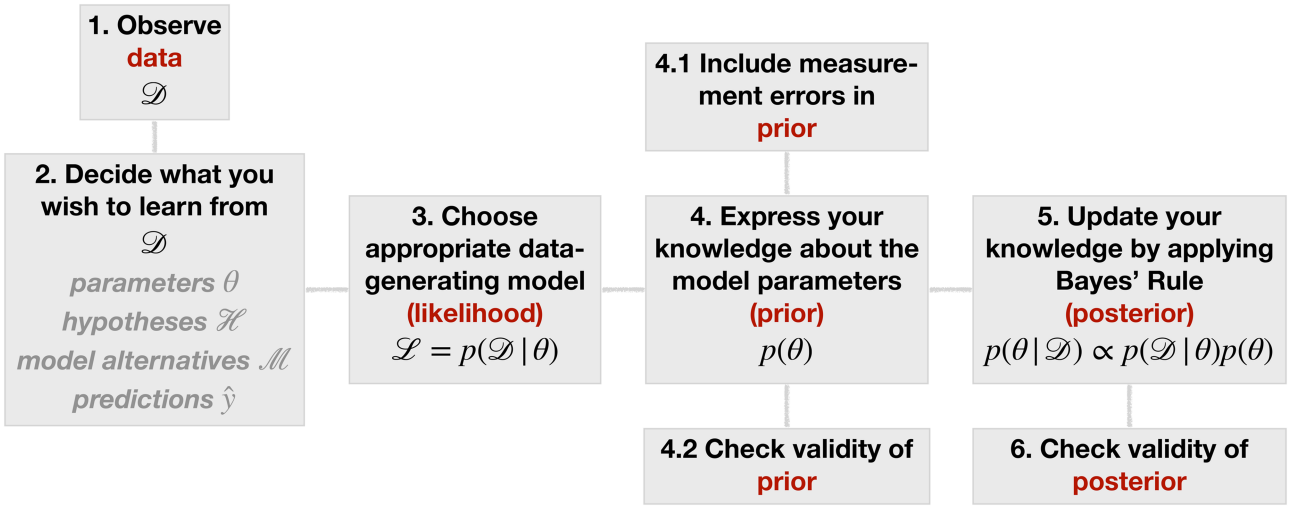

Figure 2. Schematic workflow adopted for the Bayesian model examples featured in this study. Terms in red are the main ingredients of Bayesian inference discussed in the text. Given a set of observations, we first need to decide which quantities we wish to learn from these data $\mathscr{D}$. We can learn parameters θ of a model that we assume to have generated the data (detailed here); hypotheses $\mathscr{H}$ regarding the data; model alternatives $\mathscr{M}$ or simply predictions of data $\hat{y}$ at unobserved locations. Existing information independent of the data together with assumptions about any measurement errors enter as prior knowledge, which needs thorough scrutiny before applying Bayes' rule. Similarly, the resulting posterior knowledge also needs thorough checks. [Colour figure can be viewed at wileyonlinelibrary.com]

specifying priors. In the following I present some worked examples that adopt a Bayesian viewpoint, drawing on several practical problems in geomorphology (Figure 2).

Worked Examples

Trends in landslide size: Bayesian linear regression

In simple linear regression we want to predict a continuous target variable $\mathbf{y} = \{y_1, ..., y_n\}^\mathsf{T}$ from a single, continuous predictor variable $\mathbf{x} = \{x_1, ..., x_n\}^\mathsf{T}$, for which we have n data pairs. In most computer programs we store these variables in column vectors, printed here in bold font and using a transpose symbol. We are interested in fitting a straight line of the form $y(x) = w_0 + w_1 x$ to these data. Here w_0 is the regression intercept, w_1 is the regression slope and index i identifies a single data point. Never do we encounter perfect straight line fits, so it is reasonable to assume that each observation y_i has some noise because of measurement errors or some natural variability (or both). This definition of noise, however, is contingent on the chosen model, and may exclude other aspects of data quality. In simple linear regression we assume that this noise is Gaussian with zero mean and variance σ^2. We further assume that this variance is known and fixed for all data points that we assume to be free of observation errors:

$$\begin{aligned} y_i &= w_0 + w_1 x_i + \mathcal{N}(0, \sigma^2) \\ &\text{or} \\ y_i &\sim \mathcal{N}(w_0 + w_1 x_i, \sigma^2) \end{aligned} \tag{4}$$

These equivalent formulations state that each point on the regression line is a local *mean* estimate of y_i conditional on a linear combination of x_i. The Gaussian noise term concedes that our observed measurements are sprinkled either side of the regression line, though more likely closer than farther away. You can read '$\sim$' as 'is distributed as'. The parameters that we wish to learn here are the regression coefficients. In the frequentist view, we can estimate the optimal values of w_0 and w_1 by ordinary least squares or maximum likelihood. In the Bayesian view, we can express this problem as

$$p(\mathbf{w}|\mathscr{D}) = \frac{p(\mathscr{D}|\mathbf{w})p(\mathbf{w})}{p(\mathscr{D})} \tag{5}$$

where $\mathbf{w}$ is a column vector containing the regression weights w_0 and w_1, and $\mathscr{D}$ represents all observed data pairs of x_i and y_i. The vector notation allows for easily expanding the model to multiple predictors. The likelihood of observing the target data $\mathbf{y}$ from the straight-line model with Gaussian noise (Equation (4)) thus depends on the input data $\mathbf{x}$, the regression weights $\mathbf{w}$, and the noise σ^2, and is

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i|w_0 + w_1 x_i, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(y_i - w_0 - w_1 x_i)^2} \end{aligned} \tag{6}$$

The product (the $\prod$ symbol) arises from the assumption that the data points are independently generated from the same (identical) probability distribution, a property abbreviated as i.i.d. Recall that we can multiply probabilities that are independent of each other to obtain their joint probability. The exponential term is tied to the definition of the Gaussian probability density; note how the exponent contains the squared difference between each target data point and the straight-line equation. The likelihood in Equation (6) states how plausibly any set of $\mathbf{w}$ produces the observed values $\mathbf{y}$ based on the predictor values $\mathbf{x}$ and σ^2 (Figure 3A). The global maximum of this function identifies the set of parameters that are most likely to have generated the data; hence the term *maximum likelihood*.

We specify the Bayesian model by multiplying this likelihood function with a prior distribution of $\mathbf{w}$. This prior essentially assigns weights to the regression parameters based on what we already know or believe about them without even looking at the data $\mathscr{D}$. The posterior distribution is then

$$p(\mathbf{w}|\mathbf{y}, \mathbf{x}, \sigma^2) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2)p(\mathbf{w}) \tag{7}$$

Note that this posterior is two-dimensional because we wish to learn the joint probability distribution of w_0 and w_1; recall

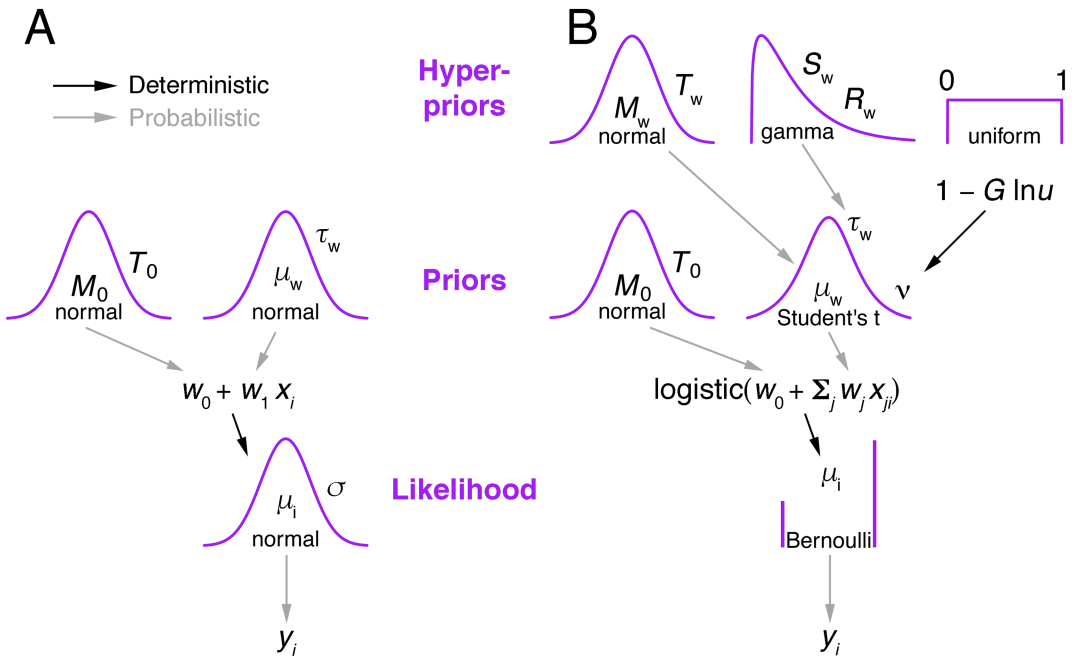

Figure 3. (A) Structure of a Bayesian simple linear regression (see text). The Gaussian distributed outcome y_i is estimated from a linear combination of an intercept w_0 and a predictor x_i weighted by w_1. Both components of this linear model have Gaussian priors. (B) Structure of a Bayesian robust multivariate logistic regression. The binary outcome y_i is estimated by a Bernoulli likelihood with a probability of success determined by a logistic linear model with intercept w_0, predictor weights w_j, and predictor values $\mathbf{x}_{ji}$. Index i refers to an individual data point, while index j can refer to either (a) different predictors or (b) one predictor with groups in the data, for example catchments or mountain ranges. The regression intercept has a Gaussian prior, and the regression weights have a robust, Student's t distributed prior with ν degrees of freedom. In a multilevel (or hierarchical) model the weight prior has three priors itself (termed *hyperpriors*) that encode how the means, spreads and ν of the weights vary across groups j. For example, data from the same catchment or mountain belt may have more similar regression parameters. Grey (black) arrows stand for 'is distributed as' ('is equal to'). Modified from Blöthe *et al.* (2019). [Colour figure can be viewed at wileyonlinelibrary.com]

that we treat σ^2 as fixed and known. It is convenient to compute the *log posterior*, because multiplying many small numbers can generate underflow errors on computers (McElreath, 2016). The log-transformed product of likelihood and prior is equal to the sum of the log likelihood and the log prior. We can interpret this sum more easily, because the maximum of the posterior distribution retains its location:

$$\ln p(\mathbf{w}|\mathbf{y}, \mathbf{x}, \beta) = -\frac{\beta}{2}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2 + \frac{n}{2}\ln\frac{\beta}{2\pi} + \ln p(\mathbf{w}) + k \quad (8)$$

Here we define the *precision* $\beta = \sigma^{-2}$ as the inverse of the variance, while k is a constant representing the log evidence and ensuring that the posterior distribution is properly normalized. We notice that the log posterior distribution depends on several terms. The first term is the sum of squared residuals that we measure as the vertical distances of each data point y_i from its model mean $y_i = w_0 + w_1 x_i$. The further the data points are away from this straight line, the less plausible are specific choices of regression weights $\mathbf{w}$, if we hold every other term in Equation (8) constant. The smaller the sum of squared residuals, the less the shape of the posterior deviates from that of the prior: the data strongly support our initial knowledge about the regression parameters. Note how the log of precision β scaled by 2π in the second term can introduce either negative or positive effects of sample size. Equation (8) also shows that many data points increasingly override the influence of the log prior.

Which distribution should we assign to $p(\mathbf{w})$ to characterize our prior knowledge about the regression parameters? In the simplest and least informed case, both w_0 and w_1 could be any real number, and we could assign equal probabilities to each. This choice would be an *improper* prior because integrating all probable outcomes over infinite bounds is infeasible. Hence a uniform probability distribution requires a lower and upper bound. If these bounds are spaced sufficiently far apart, the resulting flat prior on $\mathbf{w}$ is essentially a constant in Equation (8), and the shape of the posterior is fully determined by that of the likelihood function. The concept behind a prior, however, is that we *do* know at least something about the regression parameters.

One example of a *weakly informative prior* is a Gaussian distribution on the regression coefficients. Specifying a prior distribution $w_1 \sim \mathcal{N}(0, 1)$, for example, means that we expect the slope of the regression line to be in the interval $[-2, 2]$ with about 95% probability, or within two unit standard deviations of the zero mean. We can apply this Gaussian prior to both intercept and slope and use the same precision α for their prior distributions, hence $w_0 \sim \mathcal{N}(0, \alpha^{-1})$ and $w_1 \sim \mathcal{N}(0, \alpha^{-1})$. Keep in mind that such a weakly informed prior should ideally be replaced by more knowledge about the parameter(s), if available. To keep things simple, we assume that these two priors are independent from each other. With both the likelihood and prior being Gaussian, their product is also Gaussian (after the necessary re-normalization) and hence the log posterior is

$$\ln p(\mathbf{w}|\mathbf{y}, \mathbf{x}, \alpha, \beta) = -\frac{\beta}{2}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2 - \frac{\alpha}{2}(w_0^2 + w_1^2) + z \quad (9)$$

where z is a constant ensuring that the posterior distribution is properly normalized. We see how very low or very high values of w_0 and w_1 will penalize the posterior. The prior acts as if we added an extra data point that represents the influence of regression weights on the sum of squared residuals. Equation (9)

is the Bayesian equivalent of *ridge regression* in frequentist statistics, where it is used to penalize unduly extreme regression coefficients. We can expand on Equation (9) in many ways. For example, we could also learn the noise σ^2 (or precision β) from the data. We could consider that the noise differs for each data point. We could include measurement errors in the data by replacing single data points by distributions. In all these cases we would have to specify prior distributions for the quantities that we wish to learn. For example, if we knew initially from physical constraints that the regression coefficients are positive only and more likely close to zero than not, we could use lognormal or exponential priors on **w**.

We now apply this model to real data. In a study on 24 rock avalanches in southeast Alaska, Coe *et al*. (2018) reported that both size and mobility of these slope failures have been increasing between 1984 and 2016. Most of these landslides detached during periods of exceptionally high winter and spring air temperatures from areas modelled to have mountain permafrost. The authors excluded that any of the rock avalanches were triggered by earthquakes, but mentioned that accumulating deformation in the rock mass, glacier melt or changes in precipitation might also reduce the stability of rock slopes in the region.

Here we focus on the apparent increases in the size and mobility of rock avalanches, and use linear regression of these variables against time as a first trend analysis. We measure the size of a rock avalanche by its combined footprint of the scar, runout and deposit areas, A. The mobility of each landslide is expressed by the ratio of its total drop height over its runout, H/L. Besides these two response variables, we use the approximate time of slope failure, t, as the predictor variable. Coe *et al*. (2018) used satellite imagery to track the first occurrence of rock avalanches and were able to narrow down the failure times to several months. We use the original data and assign a timestamp t to each landslide by taking the date of the earliest image that the landslide appeared on. We assume that all errors in t are negligible compared to the study period of 32 years. If the time of each landslide has an uncertainty of half a year because of unavailable, cloudy or otherwise noisy images, the error in t would be 1.5% of the study period. Before conducting our Bayesian analysis, we standardize the data for better comparison and higher computational efficiency (Kruschke, 2015), so that target and predictor values have zero mean and unit variance. We tag standardized parameters with an asterisk.

We start by computing the log likelihood function to see how it depends on the regression intercept w_0 and slope w_1 (Equation (9); Figure 4A). The maximum (log) likelihood is identical to the mean parameter estimates that an ordinary least squares regression returns, with $w_0^* = -9.716 \times 10^{-16}$ and $w_1^* = 0.295$ (we use rounded numbers here). Following the reasoning above, our prior is a two-dimensional isotropic Gaussian distribution with a maximum at $w_0^* = 0$ and $w_1^* = 0$, and unit precision by design. This encodes that we believe that the regression parameters are independent of each other, and more likely have smaller rather than larger absolute values. The log posterior is the sum of the log likelihood and the log prior and thus combines contributions of both. We see that the log likelihood (and especially its maximum) is largely uninfluenced by our choice of prior, as the *maximum a posteriori* (MAP) estimate is very close to the maximum likelihood (Figure 4A). Yet the posterior has more closely spaced contours than the likelihood and thus attracts more probability mass or certainty towards its maximum. By definition, the posterior is a weighted compromise between the likelihood and the prior. The 24 data points are already sufficient to let the posterior inherit most of the shape of the likelihood, whereas the shape of the prior matters less. The contour values are arbitrary and need to be adjusted if re-normalizing the log posterior to its original scale.

Let us have a closer look at the posterior estimates, starting with a Bayesian linear regression of rock-avalanche area over time. Imagine a line that is parallel to the w_1^*-axis and slicing through Figure 4A. This cross-section of the posterior gives us the probability of w_1^* *conditional* on that value of w_0^* where our line intersects the w_0^*-axis. By stacking such vertical lines for all values of w_0^* and re-normalizing, we obtain the *marginal* posterior of w_1^* that is independent of w_0^*. These marginal distributions inform us about how credible the regression parameters are (Figure 4C). Yet how we summarize the posterior is up to us. Using simple point estimates such as the MAP ignores the wealth of information that a Bayesian analysis offers. Instead, we can describe the shape of the posterior using intervals. One choice is the *highest density interval* (HDI) that we obtain by slicing the posterior distribution with a horizontal line such that most of the probability mass lies above the line. We need to choose the fraction of probability mass, and thus the level of credibility of our inference. For example, the 95% HDI of the standardized intercept w_0^* is $[-0.401, 0.411]$, meaning that w_0^* is in that interval with 95% probability. This interpretation is more intuitive compared to the classic confidence interval in frequentist statistics: the 95% confidence interval contains the true parameter value in 95% of infinitely many regression experiments. In this frequentist definition 'true' and fixed parameter values generated the data at hand randomly. In the Bayesian view the data are fixed; together with our prior knowledge, they are all we have. Instead, the parameters are uncertain and expressed by probability distributions. We sequentially update these distributions as new data become available, so that the posterior of one analysis becomes the prior of the next analysis and so on.

We see that the posterior of w_0^* contains zero with 95% probability, and therefore remains ambiguous about whether the intercept is negative or positive (Figure 4C). The same applies to the standardized regression slope; while it has a posterior average of $w_1* = 0.285$, its 95% HDI contains zero and also negative values. We infer a lack in credible trend of rock-avalanche area with time, contrary to what Coe *et al*. (2018) proposed. Even ordinary least squares regression of the data returns a slope that is statistically indistinguishable from zero with $w_1^* = 0.295 \pm 0.204$ (± 1 standard error), depending on which significance level we might adopt. Retransforming the posterior regression slope to its original scale, we obtain that $-0.050 < w_1 < 0.267$ with 95% probability. Despite this ambiguity, our posterior regression weights are narrower than those from our priors. We have reduced our uncertainty about both the intercept and the slope by learning from the data.

We can now use these marginal posteriors to plot credible regression lines superimposed on the data (Figure 4D). A quick look might suggest that rock avalanches become larger with time, especially if taking into account the four large landslides since 2010. Yet we have seen that both Bayesian and classical linear regression are unsupportive of this idea, though only if fully acknowledging the associated uncertainties about the mean predictions. We could, however, narrow down the interval of posterior credibility. For example, being content with an 80% HDI, our posterior estimate would be $0.009 < w_1 < 0.210$ and thus credibly different from zero. Put differently, we would have a 20% probability of erroneously believing that average rock-avalanche size had increased over time linearly, based on the 24 cases reported and our original knowledge. Hence the choice of summarizing the posterior can be more flexible than deciding on a fixed significance level in frequentist models. Using a posterior mean, median or MAP collapses

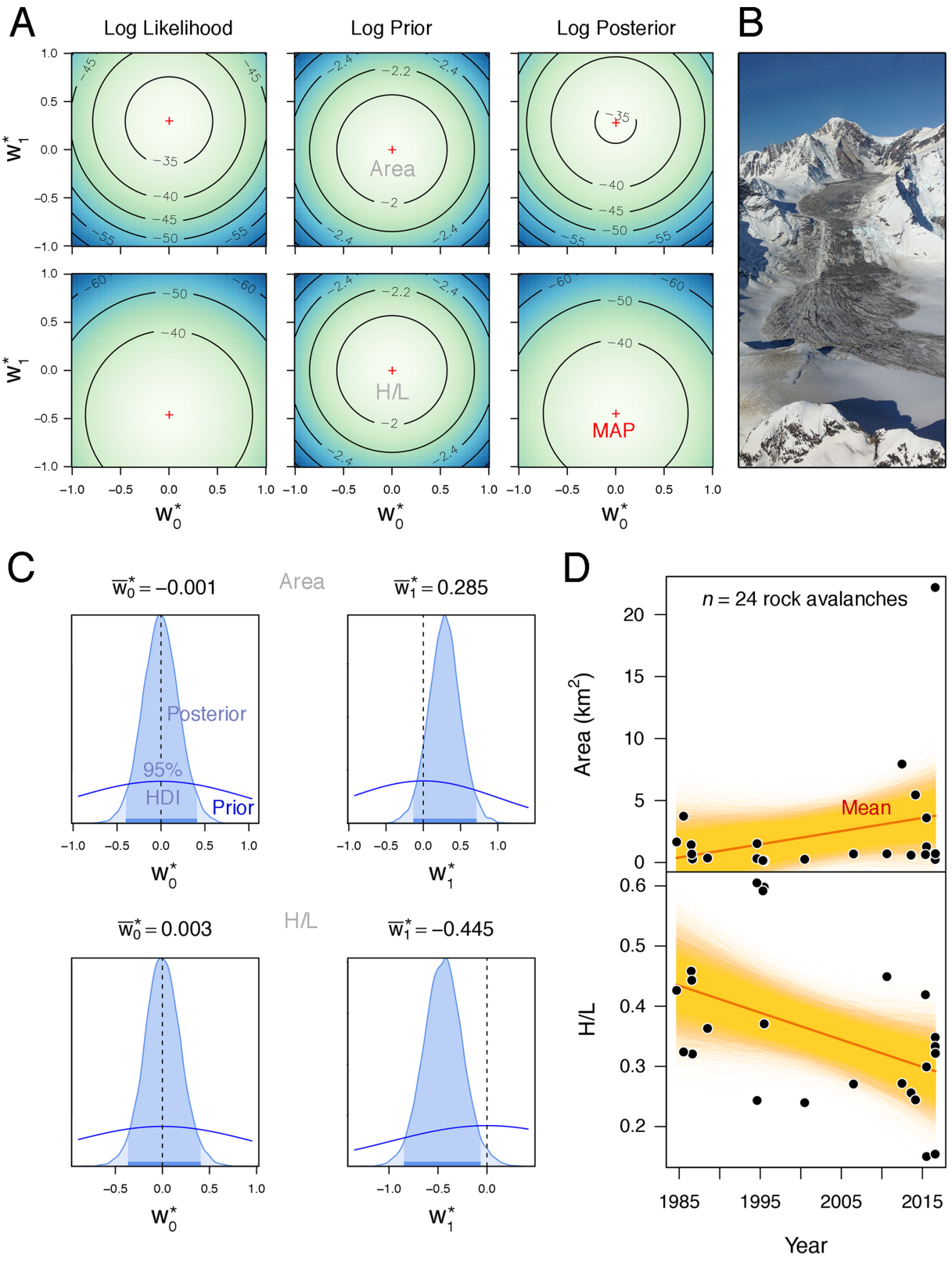

Figure 4. Bayesian linear regression of the size and mobility of 24 Alaskan rock avalanches over time; data are from Coe *et al.* (2018). (A) Contour plots of the log likelihood, log prior and log posterior as a function of the standardised intercept w_0^* and slope w_1^*. Red crosses indicate the maxima (Equation (9)), and MAP is the *maximum a posteriori* estimate; top (bottom) row refers to size (mobility). The prior precision (or inverse variance) of the Gaussian data noise is fixed at $\beta = \sigma^{-2} = 1$, and so is the precision of the standard Gaussian priors on w_0^* and slope w_1^*; that is, $\alpha = 1$. (B) The 2014 La Perouse rock avalanche, St Elias mountains, Alaska, had a total area of 5.46 km^2 and a mobility of $H/L = 0.24$. Photo courtesy of M. Geertsema. (C) Marginal posterior distributions (light blue) of the intercept w_0^* and slope w_1^*, both of which have standard Gaussian priors, $\mathcal{N}(0,1)$. The *y*-axes are omitted for clarity; it is the shape of the posterior that matters. (D) Orange lines are credible regression models based on the marginal posteriors. The red line is the linear model obtained by ordinary least squares regression. [Colour figure can be viewed at wileyonlinelibrary.com]

the information contained in the distribution to a point estimate. All these point estimates have their justification, but reveal nothing about the shape of the posterior.

In checking for a trend in rock-avalanche mobility, we see a similar, though more pronounced, effect when plotting the likelihood, prior and posterior (Figure 4C). The posterior intercept is

without a direct physical interpretation, but we obtain a credible negative posterior trend in *H/L* with a 95% HDI of [−0.0086, −0.0007] for w_1. This result supports the notion that rock avalanches in southeast Alaska have become more mobile on average over the past three decades (Coe *et al*. 2018). We learned that *H/L* of those 24 landslides decreased credibly by 0.005 per year on average in the past decades. Note the spread of credible models in both regressions: for a given failure date, the estimated average fits vary by more than a standard deviation in the response variable. The abrupt vertical cuts to the credible regression lines at the margins of Figure 4D are intentional reminders that this model is an optimal fit to the data, but a step short from a Bayesian prediction for new, unobserved data.

So far we have been learning the model parameters from the data. Yet in many cases we wish to go beyond and predict outcomes $\hat{y}_i$ for unobserved input values x_i. We achieve this by acknowledging the posterior weights of all possible values of **w**. In essence, we integrate out the posterior regression weights (and drop index *i* for clarity):

$$p(\hat{y}|\mathbf{y}, \mathbf{x}, \alpha, \beta) = \int p(\hat{y}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{y}, \mathbf{x}, \alpha, \beta)\,\mathrm{d}\mathbf{w} \qquad (10)$$

Thus we obtain for each new unobserved input a posterior *predictive distribution* of the outcome $\hat{y}$, weighted by the learned model parameters. Equation (10) states that more credible combinations of model parameters get more weight in predicting new outcomes. The first distribution under the integral is the model that presumably generates both observed and new data: in the case of simple linear regression this is the likelihood in Equation (6). The second distribution under the integral is the weight posterior that we learned from the data (Equation (5)). If both the data-generating model and the weight posterior are Gaussian, then the predictive distribution will also be Gaussian. If the prior and posterior have the same functional form as is the case here, we speak of a *conjugate prior*. Most other cases require numerical sampling to approximate the predictive distribution.

Where glaciers originate: Bayesian logistic regression

Many problems in geomorphology require us to classify (Figure 1). In the field, for example, we might ask whether a river is running through bedrock; whether a sand sheet was deposited by wind; or whether a gully is man-made. We can decide better about all these questions if we are informed by some diagnostics, perhaps fluvial potholes, well-sorted sands or historic maps. In statistical learning, this problem of deciding under uncertainty is known as classification, and the diagnostics are known as predictors. One of the most basic classifiers is logistic regression. It is part of the family of generalized linear models and therefore builds on much of the theory of linear regression outlined above. We consider here more than one predictor and hence deal with multiple logistic regression. One goal of logistic regression is to predict a bivariate outcome y_i from *m* continuous predictors that store several properties of our data. We store these data in what we call a *design matrix* **X** with each row indexing an observation, and each column indexing a predictor. We include an additional first column of ones for the regression intercept, such that the *i*th matrix row is $\mathbf{x}_i = \{1, \mathbf{x}_{i,1}, ..., \mathbf{x}_{i,m}\}$.

The outcome of y_i can be 'true' versus 'false'; 'bedrock' versus 'sediment'; 'channel' versus 'hillslope' or any other pair of mutually exclusive classes. In its basic design, logistic regression uses classes '0' and '1', and assumes that data are labelled as such and hence are free of noise. The model uses the *sigmoid function*, $\mathcal{S}(x) = (1 + \mathrm{e}^{-x})^{-1}$, to map a linear combination of predictors to the interval [0,1] that we interpret as the probability of belonging to class '1' for a given data point:

$$P(y_i = 1|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \mathrm{e}^{-w_0 - w_1 x_{i,1} - w_2 x_{i,2} - \ldots - w_m x_{i,m}}} = \frac{1}{1 + \mathrm{e}^{-\mathbf{w}^\mathrm{T}\mathbf{x}_i}} \qquad (11)$$

In analogy to Equation (4), the vector **w** holds the regression weights, while the scalar product $\mathbf{w}^\mathrm{T}\mathbf{x}_i$ is a compact notation for the linear model passed to the sigmoid function. We interpret the output of the model, $P(y_i = 1|\mathbf{x}_i, \mathbf{w})$, as the probability with which we attribute a data point $\mathbf{x}_i$ to class '1'. The probability of classifying that point as part of class '0' is the inverse probability, $P(y_i = 0|\mathbf{x}_i, \mathbf{w}) = 1 - P(y_i = 1|\mathbf{x}_i, \mathbf{w})$. Bayes' rule offers another interpretation of logistic regression. The posterior probability of a data point belonging to class '1' given the predictors $\mathbf{x_i}$ is:

$$P(y = 1|\mathbf{x}) = \frac{P(\mathbf{x}|y = 1)P(y = 1)}{P(\mathbf{x}|y = 1)P(y = 1) + P(\mathbf{x}|y = 0)P(y = 0)} \qquad (12)$$

Note that we dropped the index *i* here without loss of generality. Dividing Equation (12) by its numerator, we obtain:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \dfrac{P(\mathbf{x}|y = 0)P(y = 0)}{P(\mathbf{x}|y = 1)P(y = 1)}} = \frac{1}{1 + \mathrm{e}^{-a}} \qquad (13)$$

where

$$a = \ln\frac{P(\mathbf{x}|y = 1)P(y = 1)}{P(\mathbf{x}|y = 0)P(y = 0)} = \ln\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} \qquad (14)$$

is the *log-odds ratio*. By setting $a = \mathbf{w}^\mathrm{T}\mathbf{x}$ (and adding index *i* again), we recover the logistic regression model (Equation (11)). We see that the linear combination of weights and predictors expresses the *log* ratio of class-membership probabilities. Equation (14) shows that positive weights increase the probability of a data point belonging to class '1', whereas negative weights do the opposite. For $\mathbf{w}^\mathrm{T}\mathbf{x}_i = 0$ we obtain $P(y_i|\mathbf{x}_i) = 0.5$, meaning that both classes are equally likely for this point: it lies on the *decision boundary*. Logistic regression thus compares directly how the weight of each (standardized) predictor contributes to the probability of membership in class '1'. This is a good example of how Bayes' rule connects to a statistical inference without necessarily being 'Bayesian'. The fully Bayesian variant of logistic regression, however, needs priors on the weights as in our example of linear regression above. The main differences from the linear regression example are that the classification lacks noise in y_i and that we need to consider a different likelihood function. The bivariate outcome of y_i requires that we use a Bernoulli likelihood instead.

Logistic regression is one of the many Bayesian problems without an analytical solution. Such problems call for simplified or approximate methods, and a number of recipes are available (Caers and Hoffman, 2006). For example, Denlinger *et al*. (2012) show some of the common analytical approximations to estimate the posterior distributions applied to the heights and concentrations of volcanic ash clouds in the wake of the 2010 eruption of Eyjafjallajökull volcano,

Iceland. A widespread approach, besides analytical approximations, is to use random sampling algorithms that estimate the shape of the posterior distribution. Pioneering software such as BUGS or JAGS has been superseded by probabilistic programming languages such as STAN (https://mc-stan.org/) or Pyro (https://pyro.ai/) that cater to hierarchical models, custom prior distributions and efficient approximations of posterior distributions. If our model is simple (Figure 4), we could estimate its posterior over a dense grid in parameter space. Setting up 1000 grid points for a model intercept w_0 and a single coefficient w_1 each takes one million computations to estimate the posterior. For poorly known parameters with a wider numerical range, this choice of grid resolution might be too coarse and call for more points. Many real-world Bayesian problems have dozens of parameters, so that the computational costs quickly reach practical limits. For example, Anderson and Poland (2016) learned 25 parameters from a physical volcano model, involving quantities such as magma density, sulphur content, eruption rate and its change over time, or located deformation. Now a grid containing 1000 points for each of these parameters would require 10^{75} computations!

Hence the idea is to sample more efficiently from the posterior distributions, and several algorithms based on Markov Chain Monte Carlo (MCMC) or Hamiltonian Monte Carlo (HMC) are now implemented in many software packages. Gallagher *et al.* (2009) offered a thorough introduction to this sampling approach of solving Bayesian problems from the perspective of Earth scientists, and showcased examples of unmixing thermochronometric age distributions and sequence stratigraphy. Das *et al.* (2012) used MCMC sampling to classify landslide susceptibility along a Himalayan highway in northern India. They selected 18 predictors to classify whether a given pixel in their study area had a landslide or not. The predictors all had the same Gaussian priors with very low precision and included local slope inclination, rock type, land cover, soil depth, weathering, slope aspect and the density of streams, roads and geological lineaments. Pánek *et al.* (2016) used MCMC for a robust form of logistic regression to characterize whether several hundreds of very large ($>10^7$ m^3) landslides in the dried-out parts of the Caspian Sea basin were of terrestrial or submarine origin. Many of these giant slope failures intersect with cliffs and shorelines marking former lake levels. Both landslide volume and mobility H/L were credible predictors of whether the landslides had detached from above or below the lake level.

To illustrate how logistic regression works, we use data on Central and South Asian glaciers from the GLIMS Randolph Glacier Inventory version 6.0 (https://www.glims.org/RGI/rgi60_dl.html, last accessed 10 April 2019). This inventory contains data on the location, size and elevation of more than 95 000 glaciers between the Tien Shan and the Himalayas (Pfeffer *et al.* 2017). These glaciers become smaller and less elevated on average in a poleward direction, reflecting mainly effects of insolation and topography. Can we invert this observation? Can we infer from the size and location of a given glacier whether it originates above some specified elevation, perhaps a former equilibrium line altitude? Consider a scenario in which we are unable to correctly identify glacier source areas above this (or any other) elevation because of cloud cover or poor image quality. Glacier areas also change through time, but let us assume, for the sake of clarity, that this variability is small in our sample. Now, given the geographic latitude and area of any of these glaciers, what is the probability that it originated above 5000 m above sea level, for example?

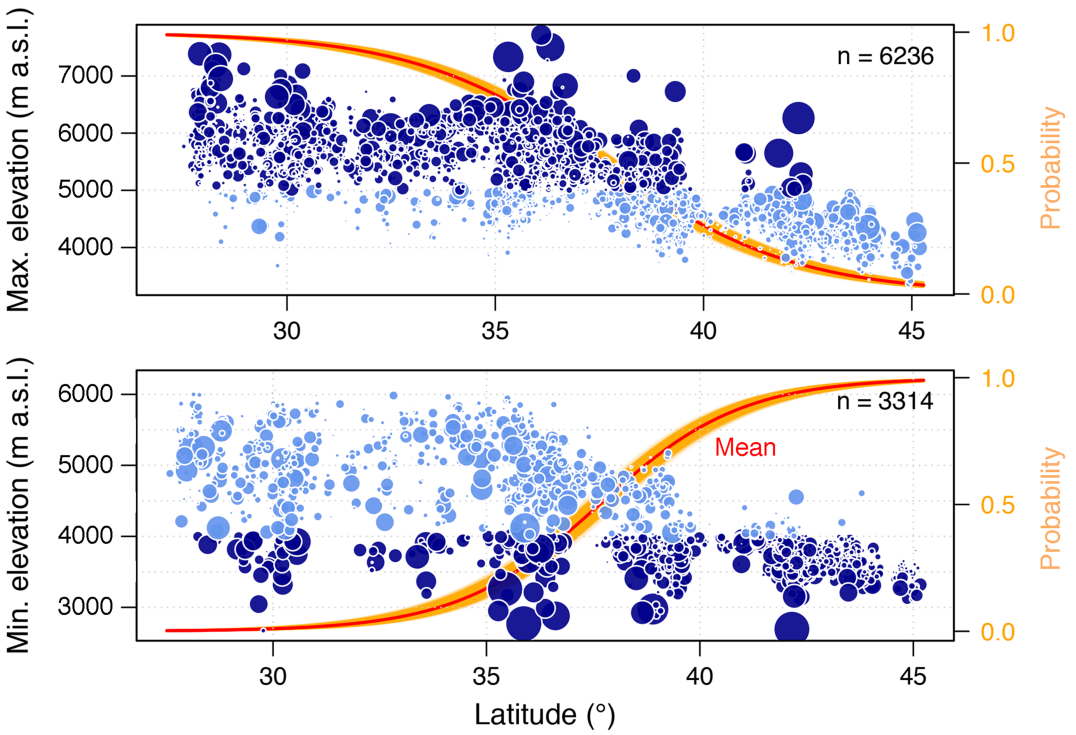

Figure 5. Bayesian logistic regression of two random samples of the maximum (top) and minimum (bottom) elevations of randomly selected Central and South Asian glaciers from the Randolph Glacier Inventory (Pfeffer *et al.* 2017) as a function of geographic latitude and glacier area (bubble size scaled to $\log_{10}$-transformed glacier area). Dark (light) blue bubbles are glaciers originating above (below) an arbitrarily chosen level of 5000 m asl in the top panel; the lower panel shows glaciers terminating above (light blue) and below (dark blue) 4000 m asl. Both classes have equal sample size. Orange (red) lines are (mean) posterior probabilities of a glacier originating above 5000 m asl (top) or terminating below 4000 m asl (bottom), given its latitude and area. For both samples, the credible decision boundaries are at about 37–38° N. [Colour figure can be viewed at wileyonlinelibrary.com]

Bayesian logistic regression can answer this question by including prior information that we encode as an isotropic Gaussian on all regression weights $\mathbf{w}$ (Figure 5). For a random sample of more than 6000 glaciers we obtain a mean posterior estimate of $\mathbf{w}^{*\mathrm{T}}\mathbf{x}_i^* = 16.522^{+0.777}_{-0.810} - 0.439^{+0.220}_{-0.200}\mathbf{x}_1^* + 0.571^{+0.117}_{-0.115}\log_{10}\mathbf{x}_2^*$, where $\mathbf{x}_1^*$ is standardized geographic latitude and $\mathbf{x}_2^*$ is standardized glacier area; the errors enclose the 95% HDI. At this level all weights are credibly different from zero, while the negative weight of latitude means that the probability of glaciers originating above 5000 m above sea level (asl) decreases in a poleward direction. The standardized regression weights tell us that geographic latitude contributes less to the classification than does glacier area on average. We can use the area under the operator receiving characteristic curve (AUROC or simply AUC) to assess how the model performed. The AUC is a widely used metric for logistic regression and summarizes the true positive rates versus the false positive rates for variable decision boundaries (Witten *et al*. 2011). The mean posterior AUC of 88.4% indicates a good performance of this classification.

Similarly, we can learn from the data the probability that a glacier of known latitude and size is terminating below a specified elevation, for example 4000 m asl. Without any other information besides latitude and glacier area, how sure can you be that a given glacier has advanced below that (or any other) elevation? Some glacier snouts might be debris covered and thus difficult to discern from the surrounding valley floor. Alternatively, assume that we observe a glacier that has advanced conspicuously far down valley, more than most others in the region. How likely is that? By estimating the posterior probability of terminating below a specified elevation, we have an objective reference. To this end, we conduct another logistic regression, using another random sample. We obtain $\mathbf{w}^{*\mathrm{T}}\mathbf{x}_i^* = -22.637^{+1.471}_{-1.431} + 0.601^{+0.038}_{-0.038}\mathbf{x}_1^* + 0.96^{+0.179}_{-0.175}\log_{10}\mathbf{x}_2^*$, which means that more poleward and larger glaciers have a higher probability of terminating below 4000 m asl (Figure 5). Again, the standardized weights show that latitude has a lower mean influence on the classification than glacier area. The posterior mean AUC of 93.6% indicates a very good classification, although this may result partly from correlated predictors. One way to avoid collinearity leading to an overconfident classification is to reduce a large set of predictors to fewer principal components. Blöthe *et al*. (2019) extracted eight principal components from 76 original candidate predictors for a robust logistic regression to predict whether rock glaciers of known size and toe elevation blocked streams in mountain ranges of Central and South Asia (Figure 3B). Similar to that study, our example demonstrates that logistic regression can help to estimate elevations of glacier snouts or their equilibrium line altitudes.

Linear and logistic regression are templates for many more sophisticated applications. A straightforward extension is to include interactions between the predictors. We can also easily introduce *basis functions* on the predictors to cater for polynomial or other nonlinear inputs: we have done so by $\log_{10}$-transforming glacier area in our logistic regression example. We can also use noise other than Gaussian; a Student *t*-distributed noise would make our regression more robust against outliers (Figure 3B). With mixed distributions we can learn from data the contributions that come from different pools. One example is to infer different sources of sediment. State-of-the-art hierarchical Bayesian modelling frameworks allow the unmixing and fingerprinting of river sediments while considering catchment-wide uncertainties of tracer materials or geochemical properties (Abban *et al*. 2016; Cooper and Krueger, 2017). We can also expand Bayesian regression models to detect distinct breaks in data trends, and the following example deals with this task.

Where channels begin: Bayesian piecewise regression

We consider a classic question in geomorphology by Montgomery and Dietrich (1988): where do channels begin? This may seem obvious in many field settings. In gridded digital elevation models, however, defining stream channel heads requires us to choose a minimum supporting drainage area to derive a stream network. This choice is arbitrary and depends on the grid resolution. Yet the location of channel heads influences many landscape metrics, including estimates of how local slope [m m^{-1}] changes with upstream contributing catchment area [km^2], and hence of channel steepness and concavity. Estimates of channel concavity, in particular, rely on the regression of slope versus catchment area (Tucker, 2004). Here we invert the problem: suppose we are interested in how reliably we can detect channel heads objectively from intentionally unlabelled measurements of slope $\mathbf{S} = \{S_1, ..., S_n\}^{\mathrm{T}}$ and their corresponding upstream catchment areas $\mathbf{A} = \{A_1, ..., A_n\}^{\mathrm{T}}$ that were taken from both hillslopes and channels. The underlying rationale is that trends of local slope versus catchment area should differ between hillslopes and channels; we avoid being biased and use randomly sampled data points from the catchment. This approach differs from one that decides beforehand on which data are parts of a hillslope or a channel and then looks for an optimal separation between the two.

Our example is Corner Creek, a 3.2 km^2 mountainous catchment on the South Wellington coast, New Zealand. We use data from a 1 m digital surface model derived from airborne LiDAR measurements in 2003 (https://www.linz.govt.nz). The regression is based on a random sample of 2000 slope-area data points to minimize effects of spatial autocorrelation that could compromise our independence assumptions about the likelihood (Figure 6A). We see the characteristic order-of-magnitude scatter and clustering in the $\log_{10}$-transformed data. We consider a piecewise regression of slope versus area that joins two linear models at a connecting change point that marks the channel head. To reduce the effect of outliers we choose a robust linear piecewise regression with likelihood

$$p(\mathbf{S}|\mathbf{A}, x_c, \mathbf{w}, \sigma^2, \nu) = \prod_{i=1}^{n} \mathcal{T}[S_i|\, w_0 + w_1 A_i + w_2 (A_i - x_c)\, I(A_i \geq x_c), \sigma^2, \nu] \quad (15)$$

where $I(\mathscr{A})$ is the *indicator function* that returns either 0 if statement $\mathscr{A}$ is false, or 1 otherwise. We define x_c as the change point (where channels begin) that links the two trend lines with slope w_1 for $A_i < x_c$, and slope w_2 for $A_i \geq x_c$. We use a Student *t*-distributed noise, $\mathcal{T}(\cdot\,|\,\cdot)$, with zero mean, scale $\sigma^2 > 0$ and $\nu = 5$ degrees of freedom. This choice means that the two trend lines that connect at x_c are much less affected by data outliers than trends assuming a Gaussian noise. It is useful to keep in mind that, for a given mean and scale, a Gaussian distribution is equivalent to a Student *t* distribution with $\nu = \infty$. Low values of ν instead put more weight on the tails of the distribution and make extreme values more likely.

We now need to specify prior distributions on all parameters of interest. Following our examples above, we assume that w_0, w_1 and w_2 are distributed as $\mathcal{N}(0, 1)$ each. We treat these priors as independent of each other and formulate them with respect

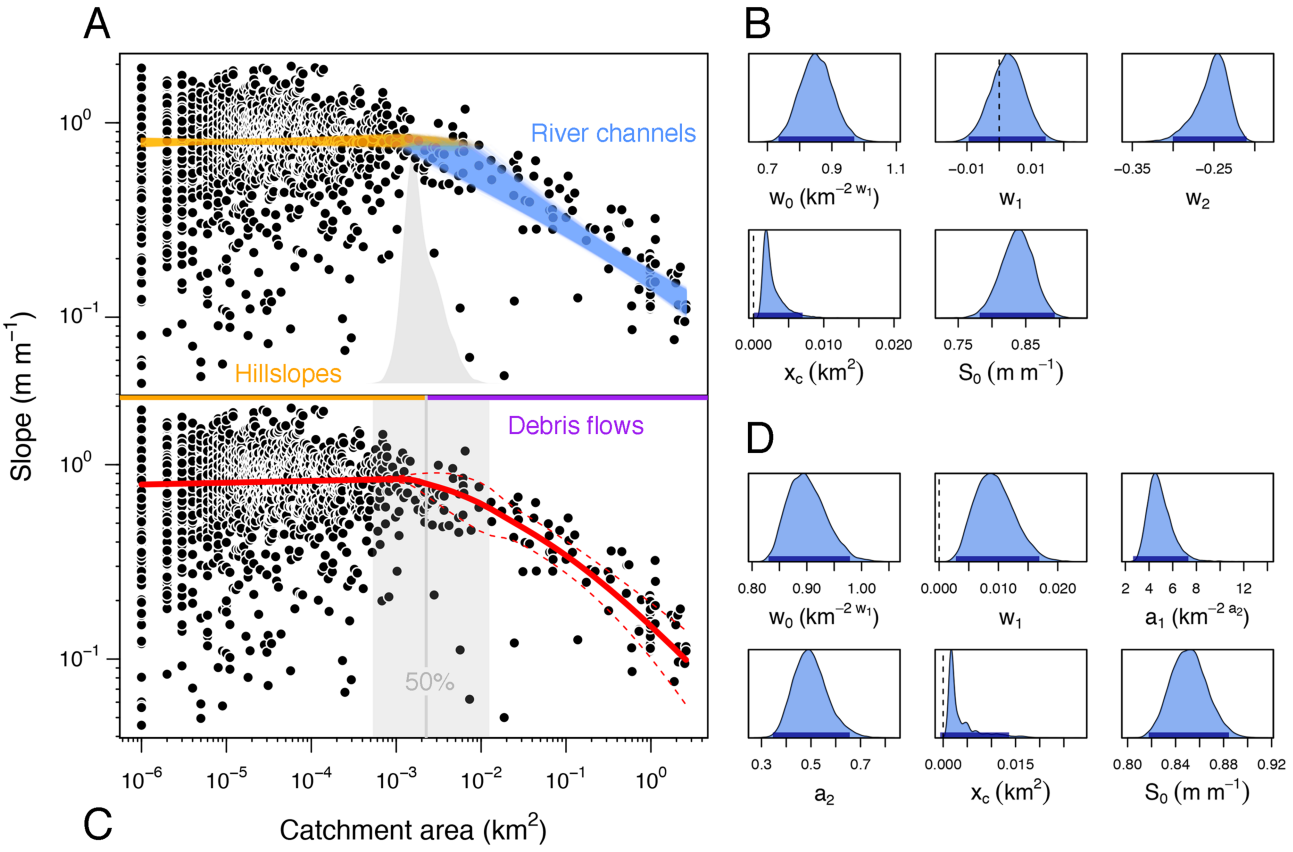

Figure 6. (A) Bayesian robust piecewise linear regression of 2000 randomly sampled slope–area data from a 1 m LiDAR model of Corner Creek, Wellington coast, New Zealand (Equation (15)). Orange (light-blue) lines are credible mean trends for hillslope (channel) grid cells. Grey probability density is the posterior of the channel head x_c linking the two trend lines. (B) Marginal posterior distributions with 95% HDIs (dark-blue horizontal bars) of intercept w_0 and slope w_1 for hillslopes, channel head x_c and the corresponding slope threshold S_0; $-w_2$ is channel concavity. (C) Alternative robust linear–curvilinear model for learning the transition from hillslopes to debris flow-dominated channels (Stock and Dietrich, 2003) (Equation (17)). The red line connects pointwise posterior means; dashed red lines are ± 2 σ about the means. The grey line (shade) is the median (95% HDI) of posterior x_c. (D) Marginal posterior distributions of model parameters as in part B; a_1 is the amplitude of the curved model segment characterizing the debris-flow domain, and a_2 is the asymptotic channel concavity for large catchment areas. Interpreting the marginal posteriors of w_0 (and a_1) makes little sense here, as their units depend on w_1 (and a_2). [Colour figure can be viewed at wileyonlinelibrary.com]

to standardized inputs of log-transformed values of **A**. We do this because slope–area plots commonly have log-transformed axes. We keep σ^2 independent of the channelization threshold, so that hillslope and channel data share the same noise. If our prior knowledge was contrary to this assumption, we would have to specify two separate priors for the two domains in slope–area space. We further assume that σ^2 is half-Cauchy distributed, following recommendations by Gelman (2006). We also need to specify our prior knowledge about the channel head x_c. We assume that x_c^* is Gaussian distributed and by choosing $\mathcal{N}(0, 0.5)$ we encode our belief that the channel head is within one standard deviation of the mean catchment area with 95% probability.

The Bayesian piecewise linear regression yields posterior trends of $w_1 = 0.003^{+0.010}_{-0.010}$ for the hillslope data, and a concavity of $-w_2 = 0.25^{+0.042}_{-0.035}$ for the channel data (units are arbitrary; Figure 6B). The nearly horizontal regression line for the hillslope data is devoid of a credible trend. This invariance of slope with respect to contributing catchment area may indicate threshold hillslopes that are inclined around a modal value. We also see that the spread of credible model slopes is much wider in the channels, owing to the fewer data. Recall that the variance σ^2 of the data with respect to the trend lines is assumed constant throughout. The posterior distribution of the change-point location shows credible channel heads that separate hillslopes from the drainage network. The channel head has a posterior mean of $x_c = 0.002^{+0.004}_{-0.001}$ km^2. This 95% HDI expresses our uncertainty about the channelization threshold given these particular 2000 data points. Similar to the logistic regression discussed above, we can assign probabilities to each DEM cell specifying whether we are more likely dealing with a hillslope or channel cell, given its upstream catchment area and posterior probability of being a channel head.

We can customize this model in several ways. For example, Stock and Dietrich (2003) suggested that slope–area data from mountainous catchments follow a curved trend instead of a power law where debris flows frequently erode into bedrock channels. They argued that debris-flow erosion might influence more than half of the length and relief of mountain river networks and proposed an empirical fit to slope–area data of the form

$$S = \frac{S_0}{1 + a_1 A^{a_2}} \tag{16}$$

where S_0 [m m^{-1}] is the slope where hillslopes give way to debris flow-dominated channels (a location Stock and Dietrich, 2003, called 'valley head'), a_1 is a coefficient [km$^{-2a_2}$], and a_2 is the asymptotic power-law exponent for large values of A and equivalent to the channel concavity in the fluvial domain. Originally, this model was meant to replace the two different trends of slope–area data for hillslope and channels by the single curve described by Equation (16). Here we keep the hillslope trend and use Equation (16) to learn where,

in terms of catchment area, hillslopes grade into debris-flow dominated channels (Nyman *et al*. 2015). We plug this equation into our Bayesian change-point model to learn the parameters from slope–area data. For the original, untransformed values of slope and area we specify likelihood functions either side of the change point:

$$\begin{aligned}\mathscr{L}_h &= p(\mathbf{S}|\mathbf{A},\, x_c,\, \mathbf{w},\, \sigma^2,\, v) \\ &= \prod_{i=1}^{n_1} \mathscr{T}(S_i|w_0 A_i^{w_1},\, \sigma^2,\, v) \text{ for} A_i \,<\, x_c \\ &\quad \text{(hillslopes)} \\ \text{and} \\ \mathscr{L}_d &= p(\mathbf{S}|\mathbf{A},\, x_c,\, \mathbf{w},\, \mathbf{a},\, \sigma^2,\, v) \\ &= \prod_{i=1}^{n_2} \mathscr{T}(S_i|\frac{w_0 x_c^{w_1}}{1+a_1 A_i^{a_2}},\, \sigma^2,\, v) \text{ for} A_i \geq x_c \text{(debris flows)}\end{aligned} \tag{17}$$

where n_1 is the number of data points in the hillslope domain and n_2 is the number of data points in the debris-flow domain; we recycle index i for each domain. The likelihood under the model is the product of the two domain-specific likelihoods, $\mathscr{L}_h\mathscr{L}_d$. By setting $S_0 = w_0 x_c^{w_1}$ we ensure that the power-law model for slope–area data of hillslopes joins with the curved relationship for the debris-flow domain. Note that this model is now nonlinear in the parameters (Equation (17)).

We encode several suitable assumptions in our priors on $\mathbf{w}$ and x_c. Without any data, we believe initially that all regression parameters w_0, w_1, a_1 and a_2 are positive. Hence the power trend for S and A in the hillslope domain is positive, and the curved trend for the debris-flow domain is convex upward. To enforce that w_0 and w_1 are positive, we use lognormal priors, assuming that $\ln w_0 \sim \mathcal{N}(0,1)$ and $\ln w_1 \sim \mathcal{N}(0,1)$. We use the data on 54 catchments outside of New Zealand (Stock and Dietrich, 2003) to inform our priors in more detail. We assign $\ln a_1 \sim \mathcal{N}(1,1)$ and $\ln a_2 \sim \mathcal{N}(-0.25, 0.25)$, as these distributions reasonably characterize these published data. Thus we explicitly and systematically use previous work to encode our prior knowledge about the regression parameters. Finally, we assume that the channel head is within the range of our catchment-area data, and distributed as $\ln x_c \sim \mathcal{N}(-1,1)$.

We obtain a posterior distribution that, in several ways, is similar to that of the simpler piecewise linear regression (Figure 6C). We learn that the posterior slope–area trend for the hillslope data is $w_1 = 0.009^{+0.007}_{-0.006}$, whereas the posterior channel concavity is $a_2 = 0.500^{+0.135}_{-0.131}$ (Figure 6D). We observe that hillslope inclinations increase minutely, but credibly, with catchment area. This positive trend reflects the choice of our lognormal prior that enforces that w_1 is positive, meaning that hillslopes become steeper towards their toes. If we believed instead that negative trends could also be plausible, we should exchange the prior for a distribution to admit also negative values. This is a good example of how the choice of the *support* (or the range of input values) of the prior distribution influences the support of the posterior distribution. From the model we also learn that channels are more concave than in the piecewise linear model. The posterior valley-head location spans more than an order of magnitude of catchment area with $x_c = 0.004^{+0.009}_{-0.003}$ km^2. We can translate this catchment position to a critical slope of $S_0 = 0.851^{+0.029}_{-0.029}$. Note that it makes little sense to interpret the marginal posterior of w_0 (and a_1), as its unit varies with the value of w_1 (and a_2). The shape of these distributions would remain, however, if we normalized $\mathbf{A}$ by a reference area, for example $A_{\text{ref}} = 1\,\text{km}^2$. Another desirable side effect of the Bayesian regression is the smooth transition between the two model segments across the uncertain valley head location. We could now use the median posterior probability of x_c to distinguish between the hillslope and debris-flow domains and put this threshold onto a map. More properly, we should compute the predictive posterior first by integrating over all possible model parameters (Equation (10)), and thus obtain a predictive distribution for each map pixel (Figure 7A).

In essence, we learned these change-point models in one single step and used slope–area data without discriminating, subjectively binning, or removing outliers. We are now able to measure our uncertainty about both channel-head locations and channel concavities in a given catchment, and move beyond the simplistic assumption of a fixed threshold of contributing catchment area by admitting and estimating the variance of this threshold (Istanbulluoglu *et al*. 2002). Capturing this variability gives us a data-driven appraisal of where channels begin; for example, we can inform the search for channel heads in the field by identifying the most likely reaches instead of single points. Moreover, we specified explicitly how well we were informed about the model originally. For example, we expected a negative sign for the slope–area trend in channel data. We also expected the locations of the channel or valley heads to be within the range of sampled data. Regardless of these or possible other prior choices, we have learned up to half a dozen model parameters (excluding the data noise σ^2) in a single and probabilistically consistent model, and avoided using incremental methods that depend on the maximized goodness-of-fit to varying subsets of the data (Stock and Dietrich, 2003). Our example model remains flexible still. We can further expand it by adding a second change point that marks the transition from the debris-flow domain to that of fluvial channels. The amount to which the posterior locations of these two change points overlap can tell us something about how credibly we can make out two distinct breaks in the slope–area data (Figure 7B). We close this example by noting that a range of diagnostics are available to compare between, and select from, these different models; examples include the widely applicable information criterion (WAIC) or leave-one-out cross-validation (LOO) (Vehtari *et al*. 2017). While this topic is beyond the scope of our example here, we have seen some ways of how we can use Bayesian analysis to learn more from existing statistical, empirical or physics-based models.

How channels widen: Bayesian Gaussian process regression

Many geomorphic problems require that we interpolate or extrapolate between a number of point measurements. Straight-line fits may be too simple and uninformative for this task. For example, we may need to create a digital elevation model from a LiDAR point cloud, derive ice-flow velocities from tracked pixels on a glacier surface, or estimate trends in channel hydraulics from surveyed cross-sections. In interpolating we assume, often tacitly, that measurements that are located closer together are more similar. 'Closer' can refer to space or time (or both) and is based on the idea that we sample from a continuum. Pick any landscape and you will find that points sufficiently close together will have similar elevations. The farther your measurements are apart, the less likely will the elevations be similar. Eventually, this correlation becomes negligible beyond a large enough distance. Hence a useful interpolation technique should take advantage of how the correlation between measured target values varies with their coordinates. If chosen suitably, this distance-dependent correlation

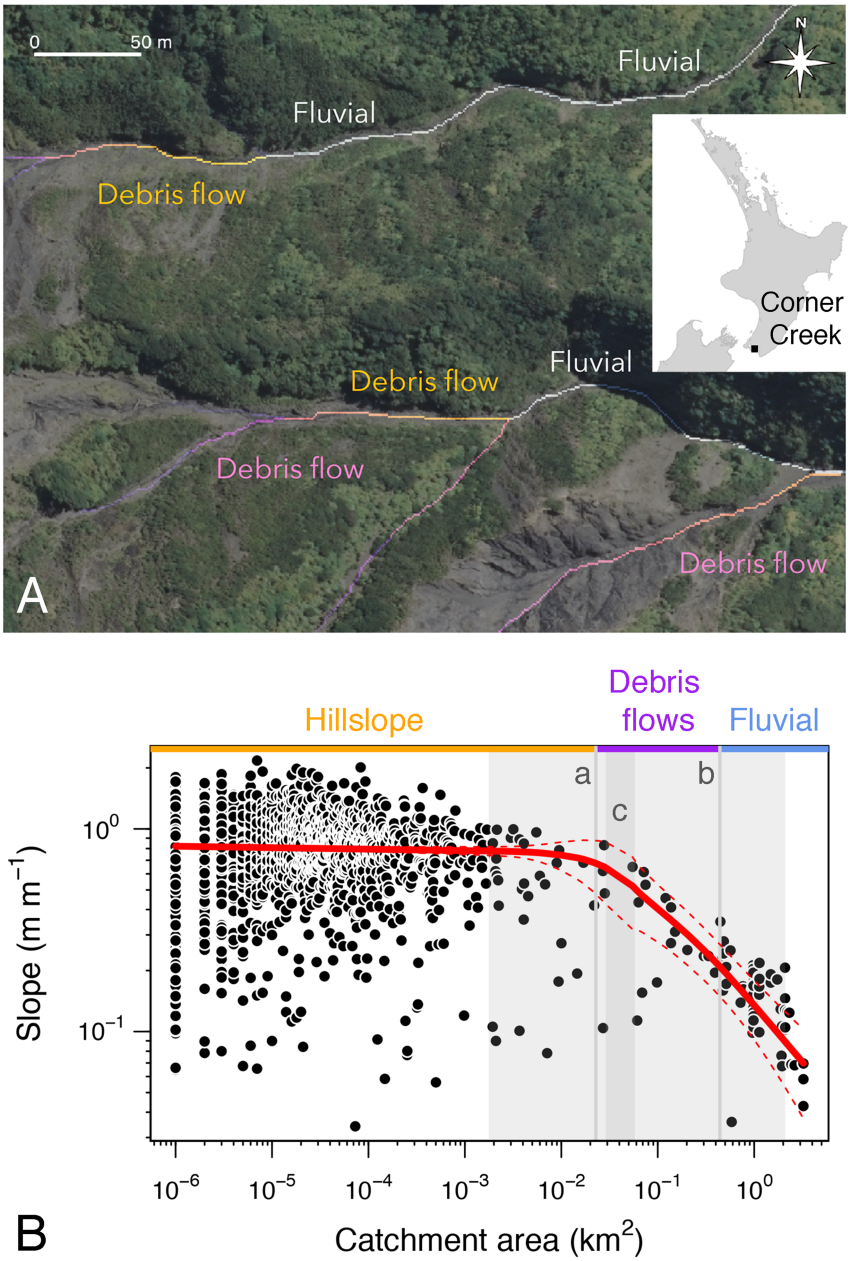

Figure 7. (A) Orthophoto detail of Corner Creek, South Wellington Coast, New Zealand (image BQ32-5K-0706; https://www.linz.govt.nz). Channel network as overlay is colour-coded with posterior probability of change-point location between debris-flow (purple to orange) and fluvial channels (white) outlined in Figure 6. (B) Bayesian robust piecewise linear regression of 2000 randomly sampled slope–area data from LiDAR-derived 1 m digital elevation model of Corner Creek, Wellington coast, New Zealand. This model is a variant of that shown in Figure 6 with two change points marking the transitions between hillslopes and debris-flow dominated channels (a) and debris-flow dominated channels and fluvial channels (b). Symbols and colours are the same as in Figure 6; note the overlap (c) of the 95% HDIs for both change-point locations. This overlap indicates that the two change points may be indistinguishable from each other for this range of catchment areas. Colour schemes in parts A and B are unrelated. [Colour figure can be viewed at wileyonlinelibrary.com]

function fits and predicts data points based on the weighted contributions of all data points. Note how we now relax the assumption of i.i.d. data by acknowledging that the data are dependent. Ignoring instead any (auto-)correlation in data can lead to severe misestimates of variances and confidence intervals (Cressie and Wikle, 2011).

We now deal with this interpolation method from a Bayesian perspective in a framework variably known as Gaussian process regression, kriging or optimal spatial prediction (Rasmussen and Williams, 2006). We can think of a *Gaussian process* as a prior distribution of mathematical functions that are of a specific type. The linear regression we used earlier is one such type of function. A prior distribution of functions means that we have infinitely many choices of functions to start with, though without being limited to straight lines (Figure 8A). By fitting a Gaussian process to data, however, we require that all these functions pass through our data points. From this constraint we obtain a posterior distribution of functions conditioned on our data. Gaussian processes are powerful, yet computationally expensive, tools for dealing with autocorrelated data (Rasmussen and Williams, 2006). The method is flexible and mathematically well explored and has many links to linear dynamic systems, stochastic differential equations, smoothing splines, Fourier transforms, Kalman filters, artificial neural networks and deep learning. More formally, a Gaussian process is a set of random variables that defines functions with input $\mathbf{x}$ and output $f(\mathbf{x})$. Note that the vector notation here refers to the number of input dimensions D, where $\mathbf{x} = \{x_1, ..., x_D\} \in \mathbb{R}^D$. We specify a Gaussian process by a mean function $m(\mathbf{x})$ and a covariance function $K(\mathbf{x},\mathbf{x}')$:

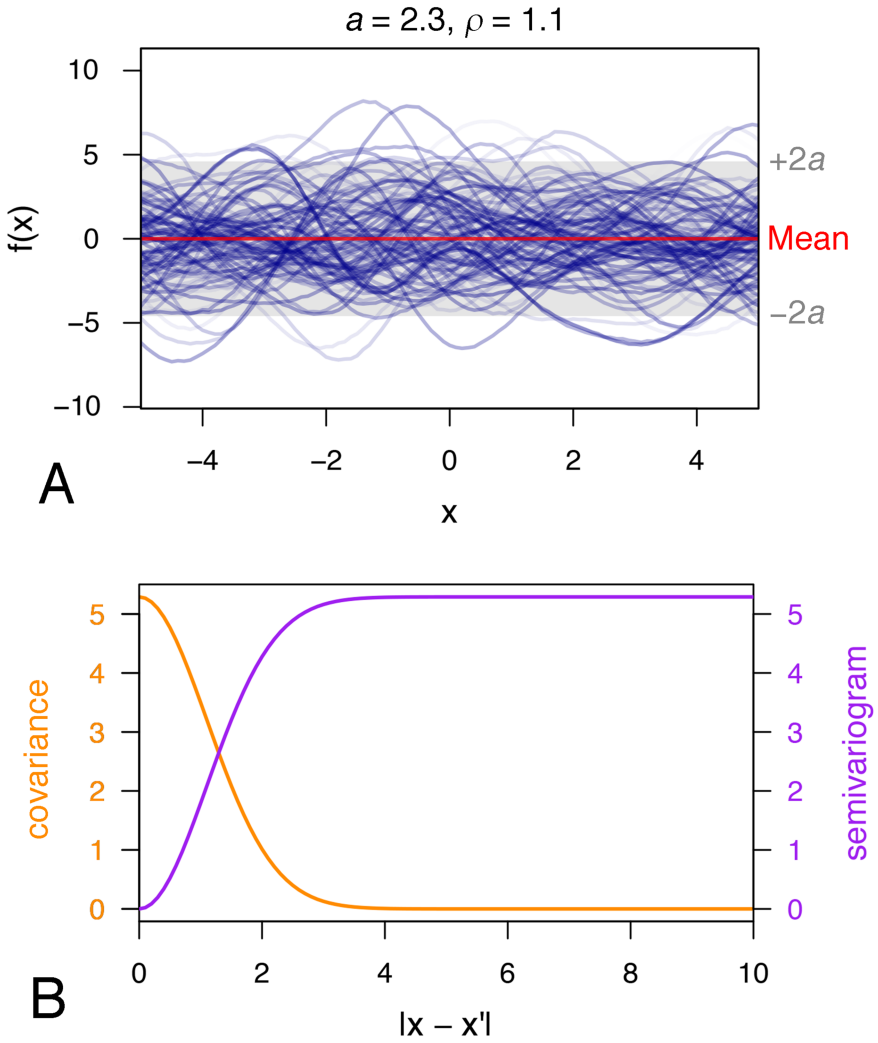

Figure 8. (A) One hundred samples (blue curves) from a Gaussian process prior with zero mean function (red line) and exponential quadratic covariance function of a single input variable x (Equation (19)). Here, $a = 2.3$ is the amplitude or marginal standard deviation of noiseless output $f(x)$ and marked by the grey shade; $\rho = 1.1$ is the length scale describing by how much we need to move along the x-axis to make the corresponding values of $f(x)$ sufficiently uncorrelated. By design, each curve oscillates around the zero mean and crosses this level upwards $(2\pi\,\rho)^{-1}$ times on average in the unit interval. Bayesian Gaussian process regression learns the posterior distribution of a and ρ such that the blue curves pass through all given data pairs $\mathscr{D} = \{x,\, f(x)\}$; these data are absent here as in any other prior. (B) Covariance and semivariogram both depend on the distance between any two input locations of each function shown in part A. Together with the zero-mean function, this single covariance function specifies, or acts a a prior over, all these functions. A local noise term σ (excluded here for clarity) adds to the contribution of a for each $f(x)$ at $|x - x'| = 0$. [Colour figure can be viewed at wileyonlinelibrary.com]

$$f(\mathbf{x}) \sim \mathscr{GP}\Big[m(\mathbf{x}),\, K(\mathbf{x},\, \mathbf{x}')\Big] \tag{18}$$

where the input values $\mathbf{x}$ and $\mathbf{x}'$ differ in their index, which can be a coordinate or timestamp, for example. The key defining characteristic is that any finite set of outputs $f(\mathbf{x})$ in Equation (18) is jointly Gaussian distributed. This means we deal with a multivariate Gaussian with mean function $m(\mathbf{x})$ and covariance function $K(\mathbf{x},\mathbf{x}')$. We can picture a Gaussian process as a multivariate Gaussian that is generalized to an infinite number of continuous random variables. In practice, we work with finite data, so that each observed data point adds a conditional constraint to the distribution. Therefore, if we condition the prior of functions that Equation (18) describes on observed data, we obtain a posterior prediction $p[\hat{f}(\mathbf{x})|f(\mathbf{x})]$ that requires the functions to pass through each data point. To avoid potentially spurious interpolation we can consider error-prone outputs by adding a Gaussian noise σ^2 as we did in the linear regression example (Equation (4)).

The functions that Gaussian processes describe are non-parametric, but we still have to decide on a mean function and a covariance function. In the most convenient case, we consider a zero mean function. This choice is useful for standardized or otherwise detrended data. However, any covariance function that we choose requires *hyperparameters*. If desired, these hyperparameters can encode the similarity of the outputs as a function of the distance between any two inputs $\mathbf{x}$ and $\mathbf{x}'$. One common textbook example is the exponential quadratic covariance function, which uses the Euclidean distance $||\mathbf{x}-\mathbf{x}'||^2$:

$$K(\mathbf{x},\, \mathbf{x}') = a^2 \exp\left[-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\rho^2}\right] \tag{19}$$

where hyperparameter a is the amplitude and hyperparameter ρ is a correlation length scale. In the case of noisy data, we would have to add σ^2 to Equation (19) at all locations where $||\mathbf{x} - \mathbf{x}'||^2 = 0$. You can visualize $K(\mathbf{x},\mathbf{x}')$ as a symmetric matrix with as many rows and columns as we have data points. Each

entry in this matrix specifies the covariance between a given pair of points. Equation (19) states that function values $f(\mathbf{x})$ are more similar if they have inputs separated by shorter distances. If the inputs are farther apart instead, we expect a lower correlation between the corresponding outputs. Many landscapes and landforms have this property, which is why we can use Gaussian processes for regression and classification problems of topographic data. We can thus interpolate at unobserved locations by obtaining a posterior distribution $p[\hat{f}(\mathbf{x})|f(\mathbf{x})]$ for each input value. Put differently, we seek to summarize the (auto-)correlation structure of the data by learning the posterior distribution of the hyperparameters. Predictions at points where data are less (more) dense will have higher (lower) uncertainties: this property is very useful for tracking how reliable our interpolation is locally.

How do we choose a meaningful covariance function and how do we specify its hyperparameters? The standard way to do this is to maximize the *marginal likelihood* to obtain the optimal hyperparameter values (Rasmussen and Williams, 2006; Beuzen *et al*. 2019). In many geostatistical studies, this procedure is part of what is known as *ordinary kriging* and built around a *semivariogram*, which is closely related to the covariance function (Figure 8B). One practical issue of traditional kriging is the need to estimate an empirical semivariogram from a sample of the data that often are assumed to be free of noise (Cressie and Wikle, 2011). These estimates can be biased, especially if the values of the hyperparameters depend on trial-and-error tuning. For likelihood functions with multiple local maxima in particular, overfitting becomes an issue for kriging methods that seek a global optimum. Yet local maxima may stand for viable alternatives or hypotheses concerning combinations of parameters that could describe the data nearly as well. A full Bayesian model of Gaussian processes instead offers a natural penalization of the marginal likelihood by specifying prior distributions over the hyperparameters. Hence the posterior distributions may highlight more than one credible way to find structure in the data.

We briefly look at Gaussian process regression to predict downstream variations in the width of river channels. How channels gradually widen in a downstream direction is an important question in estimating hydraulic geometry, flood frequency and stream power. Kriging can reveal both trends and local variations of channel geometry from limited survey data (Legleiter and Kyriakidis, 2008). The data we use here are active channel widths of Río Rayas, a medium-size gravel-bed river that drains the flanks of the Chaitén and Michinmahuida volcanoes of south-central Chile (Figure 9A). The active river bed runs through dense temperate rainforests, and raised its level following input by pyroclastic sediments from the 2008 eruption of Chaitén (Ulloa *et al*.2015). Contrasts in colour and brightness between the forest and the unvegetated river bed allow automated measurements of approximate channel width. We do this by thresholding the normalized difference vegetation index (NDVI) in a 10 m resolution Sentinel-2 image of low-flow conditions (https://scihub.copernicus.eu/dhus/#/home). Our example data are 1275 pairs of thus measured channel widths and relative downstream location along a 20 km long reach. We use 20% of these data to train our model, and the remaining 80% to estimate the prediction error.

For a Gaussian process we need to decide on a covariance function. To keep things simple, we choose the exponential quadratic covariance function introduced in Equation (19). One useful property of this particular covariance function is that, for univariate inputs x_i, the expected number of times that outputs $f(x_i)$ change from negative to positive in the unit interval is $(2\pi\ \rho)^{-1}$. You can picture this metric as an expected 'wavelength', although we deal with functions other than periodic here. For a given range of data inputs (the length of our study reach), the functions we fit and predict from can be monotone or oscillating, depending on the length scale ρ (Figure 8). Note that we earn this flexibility from a single covariance function. Now the aim is to learn from the data, and supported by some prior knowledge, the posterior distributions of hyperparameters a and ρ, and the noise σ. We assume that the Gaussian process is an acceptable model of how the mean channel width changes downstream, and add a fixed Gaussian noise σ to model local variations. Like in all Bayesian models, we need to set priors. What do we know about variations in channel width before having seen any particular data? For one, we might assume that channels become wider the more we go downstream. Yet if our study reach is too short, any trend of downstream widening might be elusive. Moreover, we might expect many local variations, where bedrock bluffs, woody debris, bank erosion or sedimentation can make the channel narrower or wider than on average. In terms of hyperparameters, we need to state what we know initially about how much active channel widths oscillate around the mean (represented by a), and how much of this is local noise (represented by σ^2). We also need to encode what we know about the length scale of changing channel widths downstream. Putting more prior weight on smaller values of ρ means that we believe that the channel narrows and widens frequently in our study reach. Putting instead more weight on high values of ρ means that we believe that much of the changes in channel width are a noisy deviation from a broader trend. The more we emphasize higher values of ρ, the more we believe in an essentially linear model of channel width versus downstream distance in our study reach. Given that we standardize our data, we choose standard Gaussian distributions for the amplitude prior and noise prior, i.e. $a \sim \mathcal{N}(0,1)$ and $\sigma^2 \sim \mathcal{N}(0,1)$, and an inverse gamma distribution on the length-scale prior, i.e. $\rho \sim \mathrm{IG}(2,1)$. The choice of inverse gamma prior for standardized ρ follows the recommendation of the Stan Development Team (2019; https://mc-stan.org/docs/2_23/stan-users-guide-2_23.pdf), and considers that ρ should exceed the minimum average spacing of measurements. Again, these choices are open to refinement according to what we know before analysing the data.

The results are instructive in several ways (Figure 9B). We see that Gaussian process regression captures and predicts much of the downstream variability of channel width, although using only one fifth of the data for training our model. The root mean squared prediction error is 16.2 m. This result is remarkable in that it hinges only on three hyperparameters (Figure 10A). Note how the uncertainties of the predicted channel widths, approximated by pointwise 95% HDIs, are highest where the density of training data is lowest. The posterior of $2\pi\rho = 805^{+105}_{-105}$ m shows what we learn about the length scale of downstream variations in channel width: we expect that the channel widens above its reach average about every 800 m. This autocorrelation aids our interpolation over shorter distances. Note that $a = 69.3^{+10.3}_{-9.5}$ m (95% HDI) quantifies the variability of the mean prediction of the Gaussian process model, whereas $\sigma = 6.2^{+1.6}_{-1.6}$ m quantifies the local noise on top of that. Beyond being able to characterize active channel width at the reach scale with these few parameters, we now also have a model to predict the posterior distribution of channel width at any location in our study reach.

We can extend this model by admitting that channel width might increase downstream in a linear manner. We thus combine a linear regression that models the downstream trend with a zero-mean Gaussian process that models the residuals. This

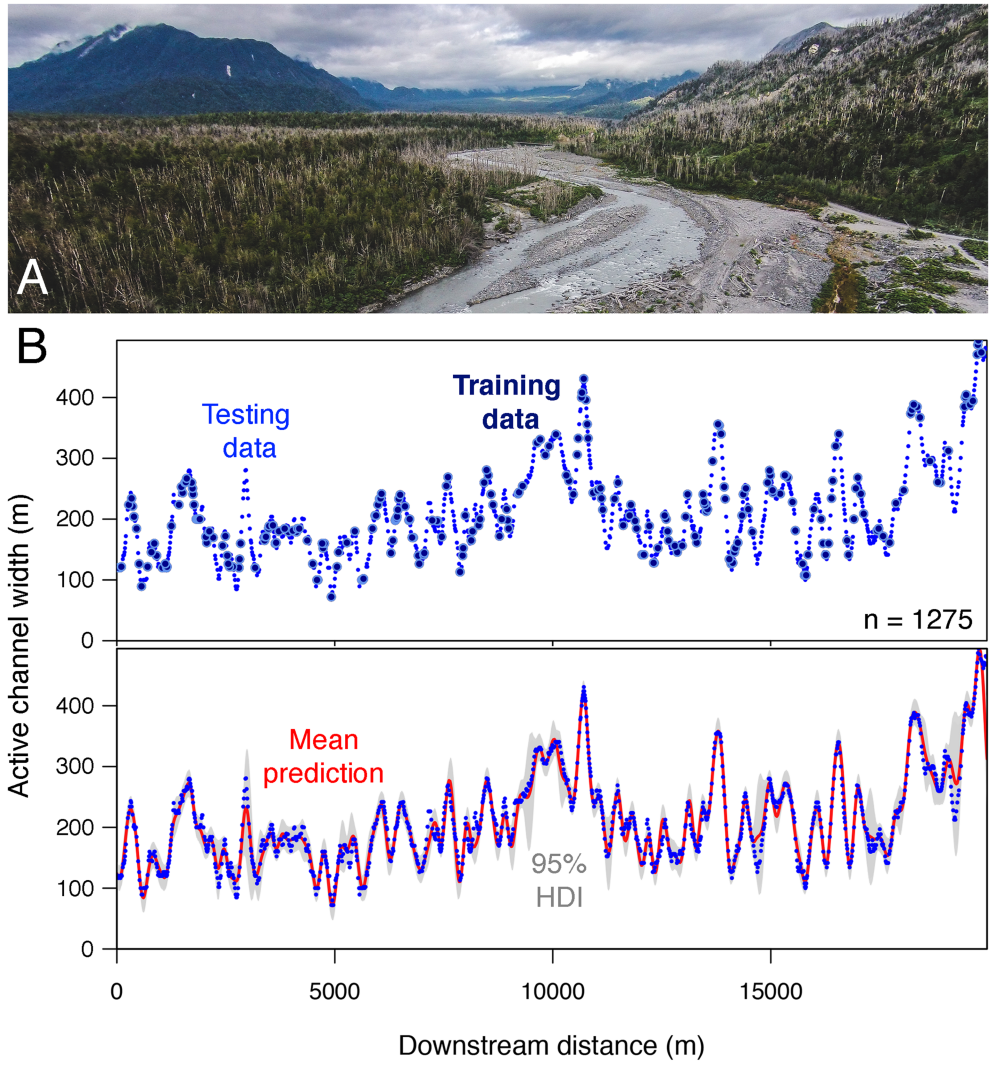

Figure 9. Bayesian prediction of channel widths of Río Rayas, south-central Chile. (A) Upstream view of active channel; this gravel-bed river is flanked by dense temperate rainforest. (B) Along-stream changes in channel width along a 20 km reach, estimated from distinct breaks in the normalized difference vegetation index (NDVI) of a Sentinel-2 image taken in April 2017. (B) Training data make up one fifth of $n = 1275$ measurements (large bubbles in upper panel); downstream distance has arbitrary origin. Posterior predictive distribution of Gaussian process regression using an exponential quadratic covariance function (lower panel; Equation (19)). The red line is the mean prediction; training data are excluded for visual clarity. [Colour figure can be viewed at wileyonlinelibrary.com]

method is also known as *universal kriging*. We can interpret this setup as a *random effects* model of local variation on top of a broader trend (Lombardo *et al.* 2019). It can be shown that this combination returns yet another, additive Gaussian process (Rasmussen and Williams, 2006). Running this model with the same priors and standard Gaussian priors on the slope and intercept of the linear model, we obtain posterior estimates of $a = 66.1^{+9.5}_{-9.1}$ m, $2\pi\rho = 757^{+90}_{-80}$ m, and $\sigma = 6.3^{+1.5}_{-1.4}$ m (Figure 10B, C). The posterior slope of the linear trend is $b_1 = 0.009^{+0.013}_{-0.009}$ and credibly different from zero judging from the 95% HDI. We learn that the active channel widens by about 9 m km^{-1} downstream on average on top of all other variations.

Two caveats deserve mention here before concluding this example. First, the exponential quadratic covariance function is very smooth and other covariance functions may reflect better the variability of channel width. A more rigorous analysis would need to compare alternative covariance functions, ideally supported by some physical reasoning. Second, the hyperparameters a and ρ are rarely identifiable as both contribute to the total variance in the data; the ratio of these hyperparameters is usually more informative. Regardless, we can use Gaussian processes as highly flexible components of models that also handle physical formulations. For example, Beuzen *et al.* (2019) coupled a Gaussian process predictor of wave runup with a morphodynamic model of coastal dune erosion in New South Wales, Australia. Fully Bayesian implementations of Gaussian processes like our example on channel width, however, are still rare in geomorphology.

Other applications

The examples presented above are building blocks that we can extend to formulate Bayesian variants of more sophisticated regression and classification problems in geomorphology. One obvious extension to these examples is to query the quality of the data. We could accommodate measurement errors in both response and predictor variables as additional prior information by replacing individual data values with distributions that express the spread around observed means (Figure 2). Our posterior would thus also include an update of our original assumptions about the data quality. In Bayesian analysis, even small sample sizes are informative. If time and cost rule out large sample numbers, for example in geochronological or thermochronometric studies, we can resort to Bayesian models to learn about variables that govern exhumation and erosion rates averaged over millions of years (Avdeev *et al.* 2011).

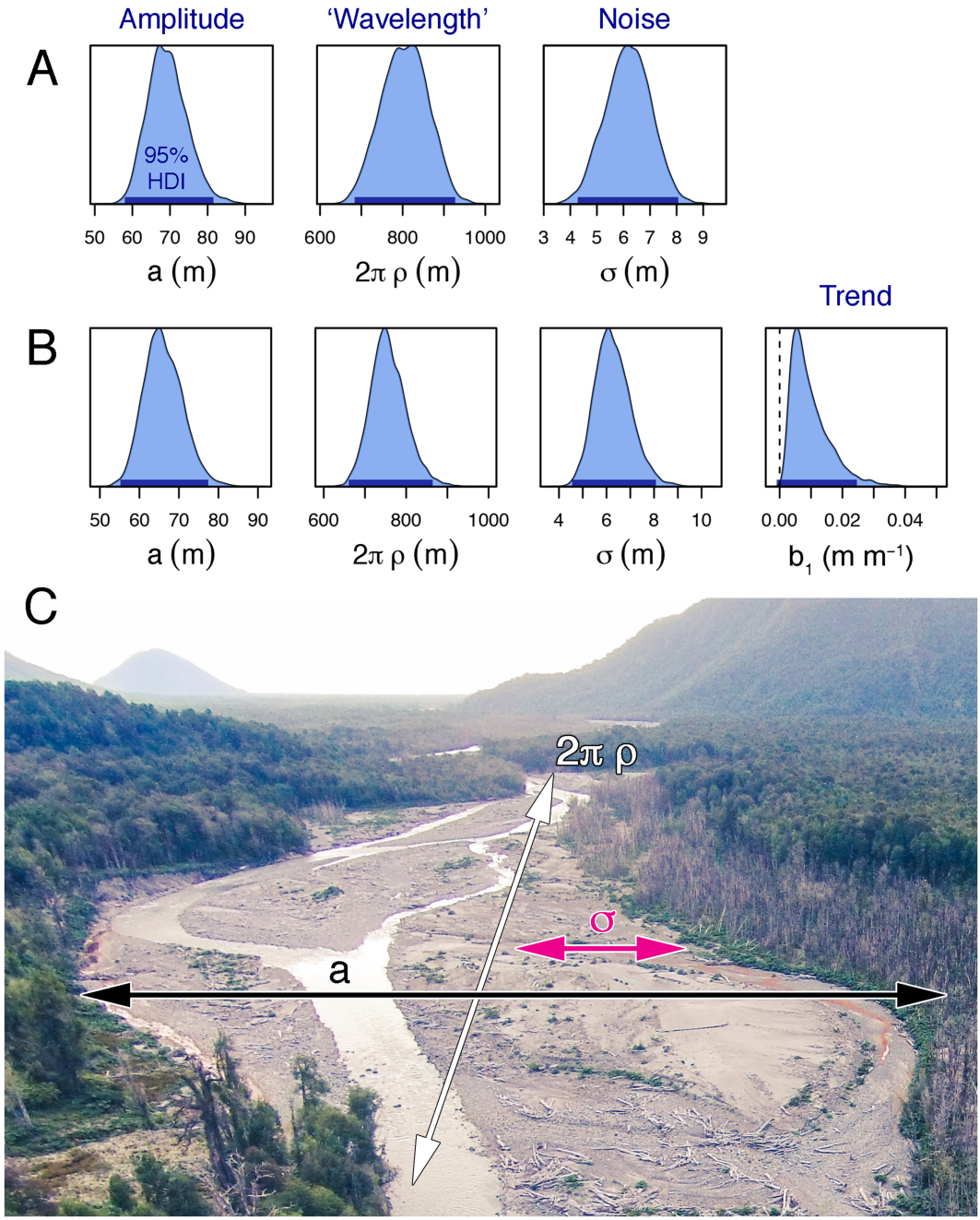

Figure 10. Posterior distributions of the parameters of two regression models using Bayesian Gaussian processes (or kriging) to predict active channel width from its downstream location, Rio Rayas, south-central Chile. (A) Bayesian ordinary kriging using squared exponential covariance function (Equation (19)); where a is the amplitude of channel width; $2\pi\rho$ is the number of upward zero crossings per unit interval or a mean 'wavelength'; and σ is an added local noise (Figure 9B). (B) Bayesian universal kriging using the same squared exponential covariance function plus a linear covariance function and the local noise term. Parameters are the same as in part A except for b_1, which is the trend of the linear model contribution estimating the downstream rate of change in channel width; dashed vertical line marks zero. In classical kriging, we only get point estimates for each of these parameters instead of entire posterior distributions. (C) Sketch highlighting how these metrics are relevant to downstream channel widening. [Colour figure can be viewed at wileyonlinelibrary.com]

Many standard statistical tests also have Bayesian counterparts in which hypothesis testing always compares alternative models, using approaches such as the Bayes factor, various information criteria, cross-validation or mixture models (Gelman *et al*. 2004; Kruschke, 2015; McElreath, 2016). The theoretical framework for some of these models was developed decades ago but now affords practical applications thanks to the available computing power that supports efficient random sampling from high-dimensional joint probability distributions (Carpenter *et al*. 2017; Bingham *et al*. 2019). Most other statistical techniques may be expanded in this way. Hence we can also use Bayesian reasoning to learn from time series data in ways that are very similar to that which we have seen for spatial data; both types of data are generated by *processes* in the statistical sense (Cressie and Wikle, 2011). In one such study, Bailer-Jones (2011) used Bayesian periodograms – plots that show how likelihood varies with frequency – to explore whether and how the rate of large (>5 km) terrestrial impact cratering had varied in the past 400 million years. Blanchet and Davison (2012) proposed a Bayesian hierarchical approach to predict from borehole data the time series of ground temperatures in mountain permafrost at various depths. Their model learns the temporal pattern of posterior ground temperature from a stochastic treatment of the heat equation and additional hyperparameters. This study showcases how to combine Bayesian inference with a widely used differential equation,

and shares several similarities with our Bayesian kriging example above. We can express the idea behind this and other hierarchical models by equipping Bayes' rule (Equation (3)) with data $\mathscr{D}$, the underlying data-generating process $\mathscr{P}$ and the parameters θ that describe this process:

$$p(\mathscr{P}, \theta|\mathscr{D}) \propto p(\mathscr{D}, \mathscr{P}, \theta) = p(\mathscr{D}|\mathscr{P}, \theta)p(\mathscr{P}|\theta)p(\theta) \qquad (20)$$

In Bayesian hierarchical models the posterior distribution is proportional to the product of a *data model* $p(\mathscr{D}|\mathscr{P}, \theta)$, a *process model* $p(\mathscr{P}|\theta)$ and a *parameter model* $p(\theta)$. In plain words, we can learn both process and parameters from the data in a single model (Cressie and Wikle, 2011).

Bayesian reasoning can also be useful when studying the recurrence of rare events. Many models in extreme-value statistics require a sufficient sample size, and can thus benefit from a Bayesian treatment if measurements are few. Extreme-value theory can be appropriate to infer the recurrence of infrequent rock avalanches as in our example above. Nolde and Joe (2013) illustrate how to learn the posterior return periods of debris flows by using a peak-over-threshold approach and eliciting expert knowledge. Their study used data on only 12 debris flows, a sample size that is prone to misestimates when fitting extreme-value distributions with classical methods. The Bayesian treatment addresses these issues by adding prior information and fully documenting uncertainty in a disciplined way. Cooley *et al*. (2006) used a Bayesian generalized extreme value distribution to model diameters of the largest lichens growing on moraine debris to infer the approximate colonization age of the substrate. Their hierarchical model included different lichen-growth curves and also spatial covariates that expressed that ages of moraines of the same glaciers covary more strongly than those of different glaciers. Bayesian inference helped Wolpert *et al*. (2016) to learn the durations of lava-dome eruptions by fitting a generalized Pareto distribution (a more flexible form of the 'inverse power law') to sample data from different volcanoes. Silva *et al*. (2015) applied a Bayesian peak-over-threshold model to estimate return periods of floods and how these might vary in response to upstream control structures and the El Niño Southern Oscillation. Similarly, Veh *et al*. (2020) inferred the 100-year peak discharge of outburst floods from moraine-dammed lakes in the Himalayas and neighbouring mountain ranges. The authors coupled a semi-empirical outburst model with data on lake size and simulated dam-breach rates. They further chose an exponential prior for the average yearly rate of outburst floods, thus emphasizing that most years were without any reported incidence.

For models with many predictor variables we can resort to *Bayesian networks* or *belief networks*, which are graph-based models of joint probability distributions (see Figure 1B,C for a most basic representation). The structure of these networks expresses conditionally dependent probabilities between these variables, which need to be discretized beforehand; we can also learn this network structure directly from the data (Vogel *et al*. 2014). Bayesian networks help in expressing visually how variables are conditioned on others and how we can substitute information if we miss out on measurements of some variables. For example, Hapke and Plant (2010) and Gutierrez *et al*. (2011) followed similar strategies of building Bayesian networks to predict yearly average trends of erosion along sections of the US coast, drawing on inputs such as local rates of sea-level rise, wave climate and coastal geomorphology. Giardino *et al*.(2019) also used a Bayesian network to model coastal erosion and beach nourishment in the Netherlands. The nodes in their network included variables such as nourishment type and volume, and rates of change in coastline position and dunes. The Bayesian network predicted the fraction of landward displacement as a function of various beach nourishment strategies. Aalders *et al*. (2011) explored a Bayesian network of some 20 factors that potentially influence peat erosion in Scotland. Most of these factors and their interactions were derived jointly during a workshop with experts. Causal links can also be encoded as conditional probabilities in Bayesian networks, and Peng and Zhang (2012) explored this property for analysing risk from dam-break floods and included various factors such as flood severity, warning times and evacuation to estimate the expected loss of lives. Essentially all the worked examples above can also be expressed as Bayesian networks, because Bayes' rule requires us to specify the joint probability distribution of all variables involved. Trees are special types of networks and offer probabilistic models for simulating sequential processes such as those that may occur during volcanic eruptions. Marzocchi *et al*. (2004) proposed a Bayesian event tree for Mount Vesuvius, Italy, to learn posterior probabilities of eruptions, their type and size, and their potential consequences for people living nearby. One advantage of trees and networks is that their structure may illustrate the relationships between variables more intuitively than the underlying mathematical formulations.

Outlook

What makes the case for encouraging more use of Bayesian reasoning in geomorphology? The learning curve can be steep and the choice of prior probability distributions may seem too arbitrary or even pointless when facing large amounts of data. In many cases, the average output of Bayesian models may seem to be hardly different from that of frequentist models (Gelman *et al*. 2004). This is reassuring, so why bother? Many concepts of frequentist statistics have a Bayesian analogue, and deciding which variant is 'better' may be pointless. However, taking a Bayesian viewpoint can be more appropriate and satisfying for many geomorphic problems. We are invited to look at these problems – and our model solutions to them – in a different light. Inversion is at the heart of Bayesian thinking and can offer instructive views of learning by combining logic and inference. The worked examples above may have motivated you to look at seeming trends in landslide size or glacier-snout elevations in a fresh way. The concept of channel heads or debris flow-dominated channels has become richer and less static by admitting variability that we fully learn from our data within a single model. Finally, we saw how to predict credible reach-scale and local variations of downstream channel widening from only a fraction of sample measurements. All that we learned about the model parameters is cast into probability distributions that allow us to predict mean channel width *and* its variability at new locations, all formal analysis of errors and their propagation conveniently included. Many other possible applications are to be charted still. If we wish to apply geomorphology to aid decisions in society, dealing with uncertainty becomes a must. Consider natural hazard and risk appraisals that are concerned with potentially adverse impacts of geomorphic processes. Such applications call for communicating uncertainties clearly (Beven *et al*. 2018a). A Bayesian approach fully and consistently treats these uncertainties, starting with our prior assumptions or knowledge and expressing uncertainty in the reproducible format of probability distributions. Posteriors are strictly conditional on observational evidence, and hence directly point at data quality and any assumptions about the data-generating process. By putting hard numbers on the unknown, Bayesian thinking encourages us to be more objective and constructive about uncertainty rather than fear it as an impediment in scientific peer review. The

Bayesian approach also obviates ad hoc methods of estimating errors and how they propagate, and offers an attractive alternative to many current black-box models in machine learning such as artificial neural networks (LeCun *et al*. 2015).

Formulating a Bayesian model forces us to spell out clearly the most important controls in a given problem. What is more, we need to acknowledge the quantities tied to these controls as intrinsically uncertain. We are required to formally encode these uncertainties as probability distributions that interact within a joint distribution. Thus one highly educational aspect of Bayes' rule is that we have to think about our model *before* seeing the data. It is a compelling way of testing and expressing objectively our prior knowledge about the quantity that we wish to learn more about before becoming possibly biased by the data. Seen this way, Bayes' rule is a good reality check that tracks each iteration in our learning process by quantifying how our prior knowledge is being refined in the light of new data. Bayesian reasoning can be challenging but also rewarding. It allows us to revise our original beliefs about a problem with the data at hand, thus encouraging us to go beyond a simple reliance on just analysing data: we need to document, and build on, what we knew before explicitly in a distribution of possible outcomes. Seen from the reverse, the new insight that the posterior provides is explicitly conditional on the data, so that we can directly measure our gain of knowledge by comparing it to our prior. Bayesian reasoning produces probabilistic predictions about the unobserved, and can be especially useful if observations are few. Simpson (2017) demonstrates this core principle of Bayes' rule in the context of (exo-)planetary geomorphology by estimating the abundance of waterworlds, using our planet's conditions as a prior.

Specifying prior knowledge in terms of a probability distribution can be challenging, or even problematic, but encourages us to document objectively what we know about a given phenomenon. Even if we know very little beforehand, we can often find a suitable prior to characterize this minute knowledge. Physical, chemical and biological problems have lower and upper bounds on the quantities that we wish to learn. We can seamlessly couple existing physical models in geomorphology with Bayesian reasoning, especially if we are interested in expressing uncertainties or if we wish to invert a given problem (Gomes *et al*. 2016; Laloy *et al*. 2017). Whether this solution is the optimal one is a different story and depends a lot on our initial choice of models and whether we think that choice is appropriate. It may turn out that mixtures or ensembles of models can be more useful than any single model, regardless of whether it has frequentist or Bayesian origins (Lavine, 2019). Either way, it seems that geomorphologists will need to understand the rapidly growing volume and diversity of data more than ever. The choice of toolkit is up to us.

Acknowledgements—Mathematical notations differ widely in texts on Bayesian methods and I have opted for some compromise. All computations were run using the **R** computational software environment (https://cran.r-project.org/) and the graphical interface *RStudio* (https://rstudio.com/). Numerical simulations of Bayesian inference were done with the STAN programming language and the RStan package (https://mc-stan.org/rstan/); the summary graphic was made with TikZ (https://github.com/pgf-tikz). Thanks are due to all developer teams. All codes are available on request. I thank Keith Beven, Lisa Luna, Christian Mohr, Georg Veh and two unidentified reviewers for helpful comments on the manuscript.

Open access funding enabled and organized by Projekt DEAL.

Conflict of interest

The author declares no competing interests.

DATA AVAILABILITY STATEMENT

The data used here are mostly from public domain resources: (1) rock-avalanche data are from Coe *et al*. (2018) (https://link.springer.com/article/10.1007/s10346-017-0879-7); (2) glacier data are from the GLIMS Randolph Glacier Inventory version 6.0 (https://www.glims.org/RGI/rgi60_dl.html, last accessed 10 April 2019); (3) LiDAR and high-resolution orthophoto data on Corner Creek are from Land Information New Zealand (https://www.linz.govt.nz/data/linz-data/elevation-dataand https://www.linz.govt.nz/data/linz-data/aerial-imagery); (4) channel-width data on Rio Rayas are derived from Sentinel-2 satellite data freely available (https://scihub.copernicus.eu/dhus/#/home) and available in processed form from the author upon request.

References

Aalders I, Hough RL, Towers W. 2011. Risk of erosion in peat soils: an investigation using Bayesian belief networks. *Soil Use and Management* **27**: 538–549.

Abban B, Thanos Papanicolaou AN, Cowles MK, Wilson CG, Abaci O, Wacha K, Schilling K, Schnoebelen D. 2016. An enhanced Bayesian fingerprinting framework for studying sediment source dynamics in intensively managed landscapes. *Water Resources Research* **52**(6): 4646–4673.

Anderson KR, Poland MP. 2016. Bayesian estimation of magma supply, storage, and eruption rates using a multiphysical volcano model: Kilauea Volcano, 2000-02012. *Earth and Planetary Science Letters* **447**(C): 161–171.

Avdeev B, Niemi NA, Clark MK. 2011. Doing more with less: Bayesian estimation of erosion models with detrital thermochronometric data. *Earth and Planetary Science Letters* **305**(3–4): 385–395.

Bailer-Jones CAL. 2011. Bayesian time series analysis of terrestrial impact cratering. *Monthly Notices of the Royal Astronomical Society* **416**(2): 1163–1180.

Barber D. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press: Cambridge, UK.

Benda L, Dunne T. 1997. Stochastic forcing of sediment routing and storage in channel networks. *Water Resources Research* **33**(12): 2865–2880.

Bennett GL, Molnar P, McArdell BW, Burlando P. 2014. A probabilistic sediment cascade model of sediment transfer in the Illgraben. *Water Resources Research* **50**(2): 1225–1244.

Berti M, Martina MLV, Franceschini S, Pignone S, Simoni A, Pizziolo M. 2012. Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach. *Journal of Geophysical Research: Earth Surface* **117**(F04006): 1–20.

Beuzen T, Goldstein EB, Splinter KD. 2019. Ensemble models from machine learning: an example of wave runup and coastal dune erosion. *Natural Hazards and Earth System Sciences* **19**(10): 2295–2309.

Beven K. 2015. What we see now: event-persistence and the predictability of hydro-eco-geomorphological systems. *Ecological Modelling* **298**: 4–15.

Beven K, Binley A. 2014. GLUE: 20 years on. *Hydrological Processes* **28**: 5897–5918.

Beven KJ, Almeida S, Aspinall WP, Bates PD, Blazkova S, Borgomeo E, Freer J, Goda K, Hall JW, Phillips JC, Simpson M, Smith PJ, Stephenson DB, Wagener T, Watson M, Wilkins KL. 2018a. Epistemic uncertainties and natural hazard risk assessment. Part 1: A review of different natural hazard areas. *Natural Hazards and Earth System Sciences* **18**(10): 2741–2768.

Beven KJ, Aspinall WP, Bates PD, Borgomeo E, Goda K, Hall JW, Page T, Phillips JC, Simpson M, Smith PJ, Wagener T, Watson M. 2018b. Epistemic uncertainties and natural hazard risk assessment. Part 2: What should constitute good practice? *Natural Hazards and Earth System Sciences* **18**(10): 2769–2783.

Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip P, Horsfall P, Goodman ND. 2019. Pyro: deep universal probabilistic programming. *Journal of Machine Learning Research* **20**: 1–6.

Blaauw M, Christen A, Bennett KD, Reimer PJ. 2018. Double the dates and go for Bayes: impacts of model choice, dating density and quality on chronologies. *Quaternary Science Reviews* **188**: 58–66.

Blanchet J, Davison AC. 2012. Statistical modelling of ground temperature in mountain permafrost. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **468**(2141): 1472–1495.

Blöthe JH, Rosenwinkel S, Höser T, Korup O. 2019. Rock–glacier dams in High Asia. *Earth Surface Processes and Landforms* **44**(3): 808–824.

Boeckli L, Brenning A, Gruber S, Noetzli J. 2012. A statistical approach to modelling permafrost distribution in the European Alps or similar mountain ranges. *The Cryosphere* **6**(1): 125–140.

Bradley DN, Tucker GE. 2013. The storage time, age, and erosion hazard of laterally accreted sediment on the floodplain of a simulated meandering river. *Journal of Geophysical Research: Earth Surface* **118**(3): 1308–1319.

Breiman L. 2001. Random forests. *Machine Learning* **45**: 5–32.

Budimir MEA, Atkinson PM, Lewis HG. 2015. A systematic review of landslide probability mapping using logistic regression. *Landslides* **12**(3): 419–436.

Burroughs SM, Tebbens SF. 2005. Power-law scaling and probabilistic forecasting of tsunami runup heights. *Pure and Applied Geophysics* **162**(2): 331–342.

Caers J, Hoffman T. 2006. The probability perturbation method: a new look at Bayesian inverse modeling. *Mathematical Geology* **38**(1): 81–100.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017. Stan: a probabilistic programming language. *Journal of Statistical Software* **76**(1): 1–32.

Charbonnier SJ, Connor CB, Connor LJ, Sheridan MF, Hernández JPO, Richardson JA. 2018. Modeling the October 2005 lahars at Panabaj (Guatemala). *Bulletin of Volcanology* **80**(4): 5–16.

Claessens L, Heuvelink GBM, Schoorl JM, Veldkamp A. 2005. DEM resolution effects on shallow landslide hazard and soil redistribution modelling. *Earth Surface Processes and Landforms* **30**(4): 461–477.

Coe JA, Bessette-Kirton EK, Geertsema M. 2018. Increasing rock-avalanche size and mobility in Glacier Bay National Park and Preserve, Alaska detected from 1984 to 2016 Landsat imagery. *Landslides* **15**: 393–407.

Cooley D, Naveau P, Jomelli V, Rabatel A, Grancher D. 2006. A Bayesian hierarchical extreme value model for lichenometry. *Environmetrics* **17**(6): 555–574.

Cooper RJ, Krueger T. 2017. An extended Bayesian sediment fingerprinting mixing model for the full Bayes treatment of geochemical uncertainties. *Hydrological Processes* **31**(10): 1900–1912.

Cressie NC, Wikle CK. 2011. *Statistics for Spatio-Temporal Data*. Wiley: Hoboken, NJ.

Das I, Stein A, Kerle N, Dadhwal VK. 2012. Landslide susceptibility mapping along road corridors in the Indian Himalayas using Bayesian logistic regression models. *Geomorphology* **179**(C): 116–125.

Denlinger RP, Pavolonis M, Sieglaff J. 2012. A robust method to forecast volcanic ash clouds. *Journal of Geophysical Research: Atmospheres* **117**(D13): 1–10.

Dey S. 2014. *Fluvial Hydrodynamics*. Springer: Berlin.

Dose V, Menzel A. 2004. Bayesian analysis of climate change impacts in phenology. *Global Change Biology* **10**(2): 259–272.

Dottori F, Di Baldassarre G, Todini E. 2013. Detailed data is welcome, but with a pinch of salt: accuracy, precision, and uncertainty in flood inundation modeling. *Water Resources Research* **49**(9): 6079–6085.

Efron B. 2013. Bayes' Theorem in the 21st Century. *Science* **340**(6137): 1177–1178.

Ferreira CM, Irish JL, Olivera F. 2014. Uncertainty in hurricane surge simulation due to land cover specification. *Journal of Geophysical Research: Oceans* **119**(3): 1812–1827.

Furbish DJ, Roering JJ, Almond P, Doane TH. 2018. Soil particle transport and mixing near a hillslope crest: 1. Particle ages and residence times. *Journal of Geophysical Research: Earth Surface* **123**(5): 1052–1077.

Gallagher K, Charvin K, Nielsen S, Sambridge M, Stephenson J. 2009. Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine and Petroleum Geology* **26**(4): 525–535.

Gelman A. 2006. Prior distributions for variance parameters inhierarchical models. *Bayesian Analysis* **1**(3): 515–533.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC: Boca Raton, FL.

George DL, Iverson RM. 2014. A depth-averaged debris-flow model that includes the effects of evolving dilatancy. II. Numerical predictions and experimental tests. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **470**(2170): 20130820.

Giardino A, Diamantidou E, Pearson S, Santinelli G, Den Heijer K. 2019. A regional application of Bayesian modeling for coastal erosion and sand nourishment management. *Water* **11**(1): 61.

Gomes GJC, Vrugt JA, Vargas EA. 2016. Toward improved prediction of the bedrock depth underneath hillslopes: Bayesian inference of the bottom-up control hypothesis using high-resolution topographic data. *Water Resources Research* **52**(4): 3085–3112.

Gutierrez BT, Plant NG, Thieler ER. 2011. A Bayesian network to predict coastal vulnerability to sea level rise. *Journal of Geophysical Research: Earth Surface* **116**(F2): 189.

Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F. 2005. Probabilistic landslide hazard assessment at the basin scale. *Geomorphology* **72**(1): 272–299.

Hapke C, Plant N. 2010. Predicting coastal cliff erosion using a Bayesian probabilistic model. *Marine Geology* **278**(1–4): 140–149.

Hassall KL, Dailey G, Zawadzka J, Milne AE, Harris JA, Corstanje R, Whitmore AP. 2019. Facilitating the elicitation of beliefs for use in Bayesian belief modelling. *Environmental Modelling and Software* **122**(104539): 1–9.

Heiser M, Scheidl C, Eisl J, Spangl B, Hübl. 2015. Process type identification in torrential catchments in the eastern Alps. *Geomorphology* **232**: 239–247.

Istanbulluoglu E, Tarboton DG, Pack RT, Luce C. 2002. A probabilistic approach for channel initiation. *Water Resources Research* **38**(12): 61–1–61-14.

Jensen JL, Hart JD, Willis BJ. 2006. Evaluating proportions of undetected geological events in the case of erroneous identifications. *Mathematical Geology* **38**(2): 103–112.

Jo S, Kim G, Jeon J-J. 2016. Bayesian analysis to detect abrupt changes in extreme hydrological processes. *Journal of Hydrology* **538**(C): 63–70.

Kadane JB. 2011. *Principles of Uncertainty*. Chapman & Hall: New York.

Kern AN, Addison P, Oommen T, Salazar SE, Coffman RA. 2017. Machine learning based predictive modeling of debris flow probability following wildfire in the intermountain western United States. *Mathematical Geoscience* **49**: 717–735.

Kruschke JK. 2015. *Doing Bayesian Data Analysis*. Academic Press: San Diego, CA.

Laloy E, Beerten K, Vanacker V, Christl M, Rogiers B, Wouters L. 2017. Bayesian inversion of a CRN depth profile to infer Quaternary erosion of the northwestern Campine Plateau (NE Belgium). *Earth Surface Dynamics* **5**(3): 331–345.

Lavine M. 2019. Frequentist, Bayes, or other? *The American Statistician* **73**: 312–318.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521**: 436–444.

Legleiter CJ, Kyriakidis PC. 2008. Spatial prediction of river channel topography by kriging. *Earth Surface Processes and Landforms* **33**: 841–867.

Lombardo L, Bakka H, Tanyas H, Westen C, Mai PM, Huser R. 2019. Geostatistical modeling to capture seismic-shaking patterns from earthquake-induced landslides. *Journal of Geophysical Research: Earth Surface* **124**(7): 1958–1980.

Marzocchi W, Sandri L, Gasparini P, Newhall C, Boschi E. 2004. Quantifying probabilities of volcanic events: The example of volcanic hazard at Mount Vesuvius. *Journal of Geophysical Research: Solid Earth* **109**: B11201, 1–18.

McElreath R. 2016. *Statistical Rethinking*. CRC Press: Boca Raton, FL.

Miller DJ, Burnett KM. 2008. A probabilistic model of debris-flow delivery to stream channels, demonstrated for the Coast Range of Oregon, USA. *Geomorphology* **94**(1–2): 184–205.

Molnar P, Anderson RS, Kier G, Rose J. 2006. Relationships among probability distributions of stream discharges in floods, climate, bed

load transport, and river incision. *Journal of Geophysical Research: Solid Earth* **111**(F2): 1–10.
Mondini AC, Marchesini I, Rossi M, Chang K, Pasquariello G, Guzzetti F. 2013. Bayesian framework for mapping and classifying shallow landslides exploiting remote sensing and topographic data. *Geomorphology* **201**(C): 135–147.
Montgomery DR, Dietrich WE. 1988. Where do channels begin? *Nature* **336**: 232–234.
Nolde N, Joe H. 2013. A Bayesian extreme value analysis of debris flows. *Water Resources Research* **49**(10): 7009–7022.
Nyman P, Smith HG, Sherwin CB, Langhans C, Lane PNJ, Sheridan GJ. 2015. Predicting sediment delivery from debris flows after wildfire. *Geomorphology* **250**(C): 173–186.
O'Hagan A, Oakley JE. 2004. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering and System Safety* **85**(1–3): 239–248.
Pánek T, Korup O, Minár J, Hradecký J. 2016. Giant landslides and highstands of the Caspian Sea. *Geology* **44**(11): 939–942.
Peng M, Zhang LM. 2012. Analysis of human risks due to dam-break floods. Part 1: A new model based on Bayesian networks. *Natural Hazards* **64**(1): 903–933.
Pfeffer WT, Arendt AA, Bliss A, Bolch T, Cogley JG, Gardner AS, Hagen J-O, Hock R, Kaser G, Kienholz C, Miles ES, Moholdt G, Mölg N, Paul F, Radic V, Rastner P, Raup BH, Rich J, Sharp MJ, The Randolph Consortium. 2017. The Randolph Glacier Inventory: a globally complete inventory of glaciers. *Journal of Glaciology* **60**(221): 537–552.
Raffaele L, Bruno L, Wiggs GFS. 2018. Uncertainty propagation in aeolian processes: from threshold shear velocity to sand transport rate. *Geomorphology* **301**: 28–38.
Ramsey CB. 2009. Bayesian Analysis of Radiocarbon Dates. *Radiocarbon* **51**: 337–360.
Rasmussen CE, Williams CKI. 2006. *Gaussian Processes for Machine Learning*. MIT Press: Cambridge, MA.
Regmi NR, Giardino JR, Vitek JD. 2010. Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA. *Geomorphology* **115**(1–2): 172–187.
Rustomji P, Wilkinson SN. 2008. Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. *Water Resources Research* **44**(9): W09435.
Schwanghart W, Worni R, Huggel C, Stoffel M, Korup O. 2016. Uncertainty in the Himalayan energy–water nexus: estimating regional exposure to glacial lake outburst floods. *Environmental Research Letters* **11**(074005): 1–9.
Seidou O, Ouarda TBMJ, Barbet M, Breneau P, Bobée P. 2006. A parametric Bayesian combination of local and regional information in flood frequency analysis. *Water Resources Research* **42**: 1–21.
Shikakura Y. 2014. Marine terraces caused by fast steady uplift and small coseismic uplift and the time-predictable model: case of Kikai Island, Ryukyu Islands, Japan. *Earth and Planetary Science Letters* **404**(C): 232–237.
Silva AT, Portela MM, Naghettini M, Fernandes W. 2015. A Bayesian peaks-over-threshold analysis of floods in the Itajaí-açu River under stationarity and nonstationarity. *Stochastic Environmental Research and Risk Assessment* **31**(1): 185–204.
Simpson F. 2017. Bayesian evidence for the prevalence of waterworlds. *Monthly Notices of the Royal Astronomical Society* **468**(3): 2803–2815.
Stefanescu ER, Bursik M, Cordoba G, Dalbey K, Jones MD, Patra AK, Pieri DC, Pitman EB, Sheridan MF. 2012. Digital elevation model uncertainty and hazard analysis using a geophysical flow model. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **468**(2142): 1543–1563.
Stock J, Dietrich WE. 2003. Valley incision by debris flows: evidence of a topographic signature. *Water Resources Research* **39**(4): 1089, 1–25.
Strenk PM, Wartman J. 2011. Uncertainty in seismic slope deformation model predictions. *Engineering Geology* **122**(1–2): 61–72.
Tucker GE. 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.
Turowski JM, Hodge R. 2017. A probabilistic framework for the cover effect in bedrock erosion. *Earth Surface Dynamics* **5**(2): 311–330.
Ulloa H, Iroumé A, Picco L, Korup O, Lenzi MA, Mao L, Ravazzolo D. 2015. Massive biomass flushing despite modest channel response in the Rayas River following the 2008 eruption of Chaitén volcano, Chile. *Geomorphology* **250**(C): 397–406.
Veh G, Korup O, Walz A. 2020. Hazard from Himalayan glacier lake outburst floods. *Proceedings of the National Academy of Sciences of the USA* **117**(2): 907–912.
Vehtari A, Gelman A, Gabry J. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**: 1413–1432.
Vogel K, Riggelsen C, Korup O, Scherbaum F. 2014. Bayesian network learning for natural hazard analyses. *Natural Hazards and Earth System Science* **14**(9): 2605–2626.
Wheaton JM, Brasington J, Darby SE, Sear DA. 2010. Accounting for uncertainty in DEMs from repeat topographic surveys: improved sediment budgets. *Earth Surface Processes and Landforms* **35**: 136–156.
Witten IH, Frank E, Hall MA. 2011. *Data Mining*. Morgan Kaufmann: Burlington, MA.
Wolpert RL, Ogburn SE, Calder ES. 2016. The longevity of lava dome eruptions. *Journal of Geophysical Research: Solid Earth* **121**: 676–686.
Zadeh LA. 2006. Generalized theory of uncertainty (GTU)–principal concepts and ideas. *Computational Statistics and Data Analysis* **51**(1): 15–46.

