AGU | ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

Key Points:
- Blanket peat restoration interventions significantly reduce flood risk to the downstream communities at risk
- Revegetation and gully blocking of eroded blanket peat reduces flood risk primarily through increased surface roughness and not storage
- Impact of gully blocking in reducing increasing "kinematic" surface storage is very significant and equivalent to revegetation alone

Supporting Information:
Supporting Information may be found in the online version of this article.

Correspondence to:
S. Goudarzi,
salim.goudarzi@newcastle.ac.uk

Blanket Peat Restoration: Numerical Study of the Underlying Processes Delivering Natural Flood Management Benefits

Salim Goudarzi[1] , David G. Milledge[1] , Joseph Holden[2] , Martin G. Evans[3] , Timothy E. H. Allott[3] , Emma L. Shuttleworth[3] , Michael Pilkington[4], and Jonathan Walker[5]

[1]Department of Civil Engineering, Newcastle University, Newcastle, UK, [2]Water@leeds, School of Geography, University of Leeds, Leeds, UK, [3]Department of Geography, University of Manchester, Manchester, UK, [4]Moors for the Future Partnership, Fieldhead, UK, [5]Department of Biosciences, Swansea University, Swansea, UK

Abstract Restoration of eroded blanket peatlands through revegetation and gully blocking is observed to also deliver significant natural flood management (NFM) benefits (reduce and delay floodpeaks). But there is a lack of clear understanding regarding how different catchment processes interact/counteract under each intervention scenario. We seek to provide more insight by rigorously calibrating TOPMODEL rainfall-runoff model to different experimental catchments each representing an intervention scenario. Through numerical experimentation with the calibrated parameters, we estimate the impact-magnitude of different processes. Our findings confirm the NFM benefits of these restoration-focused interventions. In both interventions and in our largest storms, both the delay and reduced floodpeaks are primarily due to surface roughness reducing the floodwave speed thus thickening the overland flow; we conceptualize this as an increase in a "kinematic storage." Impact of gully blocking in increasing kinematic storage is very significant and comparable to that of revegetation alone. Interventions' impact on "static storage" (interception + ponding + evapotranspiration) becomes important for smaller storms. Although interventions always increase lag times, they can be less effective in reducing peak magnitude when maximum rainfall intensity is sustained for durations longer than mean catchment delay. We propose two approaches to further increase catchment's static and kinematic storage. Finally, while our field-scale numerical study contributes to the evidence-base for NFM's effectiveness, it also provides a basis for modeling these interventions in the future. Such catchment-scale numerical studies are necessary to extend our findings to spatial scales where flooding can cause socioeconomic damage, and to provide a tool for optimizing the distributed configuration of these interventions.

1. Introduction

Peatlands cover only 2.84% of the land surface of the planet, yet they provide the human population with a wide range of services, including more than a third of soil carbon storage, river flow regulation, and biodiversity (Xu et al., 2018). These sensitive systems require high precipitation or impeded drainage to develop and remain stable. Industrialization, mechanization, land-use change (e.g., agriculture or forestry) and extraction of peat for horticultural and energy production have all led to degradation of peatlands.

Blanket peat, the dominant UK peat type, is unusual among peatlands in that it occurs on slopes up to 15° (Lindsay et al., 1988). When surface vegetation is damaged, their sloping nature means that erosion can rapidly occur (Evans, Allott, et al., 2005; Evans, Warburton, & Yang, 2006). Rapid and extensive peatland degradation has been going on in the UK uplands for at least a century due to increased atmospheric pollution (e.g., Rothwell et al., 2007), drainage, grazing, and wildfire damage (e.g., Evans & Warburton, 2011). Peatland degradation not only poses a severe threat to an important ecosystem, but in some circumstances may also enhance downstream flood risk (Acreman & Holden, 2013). For peat to develop and persist it must be largely saturated for much of the year, thus even intact blanket peatlands are often able to store little water during rainfall events, relative to other landscapes. Blanket peatlands typically exhibit a flashy response to rainfall, dominated by saturation-excess overland flow and/or near surface throughflow with minimal base-flow contribution (Evans, Burt, et al., 1999). Peatland degradation through, for instance, vegetation

loss and gully erosion can further increase the flashiness of the runoff response and increase flood risk downstream by raising peak discharge and reducing lag times (Grayson & Rose, 2010).

In the UK, a growing recognition of the value of peatland ecosystem services has led to a proliferation of programs to prevent further degradation and ultimately to restore their natural functions. In parallel, UK flood risk management policy and practice is increasingly interested in the potential for natural flood management (NFM) – interventions to "reduce flood hazard, while also sustaining or enhancing other potentially significant co-benefits including enhanced ecosystem services" (Dadson et al., 2017). Given that many river systems in the UK originate in upland catchments with a blanket peat cover, peatland restoration activities have the potential to offer significant NFM benefit. These activities often involve: drain blocking, gully blocking and re-profiling, bare peat stabilization and revegetation (Parry et al., 2014). In the UK alone, over the past two decades, extensive peatland restoration programs have been implemented (Armstrong et al., 2010; Evans, Allott, et al., 2005; Parry et al., 2014; Wallage et al., 2006). Although the impact of artificial drainage on peatland hydrology has been studied extensively, for example, Holden, Chapman, and Labadz (2004) studies of the effect of restoration efforts have primarily focused on sediment dynamics (e.g., Shuttleworth, Evans, Hutchinson, & Rothwell, 2015), carbon release (e.g., Dixon et al., 2014; Holden, 2005), and vegetation recovery (e.g., Cole et al., 2014). In fact, there is very little experimental evidence available on the effects of restoration activities on hydrological functions of peatlands.

Most recently, in the first controlled field-scale experiment on the impact of revegetation and gully blocking on hydrological functions, Shuttleworth, Evans, Pilkington, et al. (2019) published the results of their Before-After-Control-Intervention (BACI) study, showing that: lag times increased by 106% and magnitude of flood peaks decreased by 27% after revegetation, and that lag times increased by 200% and magnitude of flood peaks decreased by 51% after a combined treatment of revegetation and gully blocking. For more information regarding these values and how they are obtained, refer to Shuttleworth, Evans, Pilkington, et al. (2019). Such BACI-type experimental data enable a change in discharge characteristics to be linked to an intervention in a particular catchment, while eliminating the noise due to "natural temporal variability" resulting from inter-annual variation in synoptic hydrometeorology. These observations can later be assigned to a number of factors that are conceivable from a physical point of view (see Shuttleworth, Evans, Pilkington, et al., 2019).

However, solely based on variations in measured rainfall-discharge relationship, it is difficult to identify which hydrological processes are driving the observed changes. For example, it is not obvious whether the additional interception storage volume created by revegetating the bare peat areas can be sufficient to account for some of the observed peak magnitude reduction, and whether the time it takes to fill such additional storage accounts for some of the observed peak timing delay; similarly, for the additional in-channel, pond-storage behind the blocks. It is also not clear how important is the reduction in the response-speed (celerity) of overland flow, that results from flow on a surface made rougher by more vegetation cover and/or stream blocking. In relation to evapotranspiration, it is not known how important a role is played by increased evapotranspiration rates resulting from a larger vegetated catchment area, or additional direct pond evaporation in the gully blocked case. More importantly, perhaps, how do various processes interact (or counteract)? In reality all of the above processes, at least to some degree contribute to the observed changes to the hydrograph behavior following interventions. Thus, studying both the "absolute" and the "relative" (to one another) magnitude of impact of these processes helps better understand the underlying mechanism(s) through which each intervention primarily delivers NFM benefit (i.e., reduce peak magnitudes and increase lag times). This is desirable because it would: (1) condition our expectations for future interventions; (2) guide future decision making and implementation.

To this end, we invert the TOPMODEL rainfall-runoff model to relate the BACI observations to more specific hydrological processes by attempting to link the response to intervention in each catchment to numerical model parameters through rigorous calibration. We then use the calibrated parameter-sets to perform a series of numerical experiments to see, through the lens of a numerical model, which underlying process(es) may be driving the observed changes. In doing so, we estimate and analyze the "relative" and "absolute" magnitude of impact of each process. Finally, we discuss the implications of our findings regarding NFM at catchment-scale.
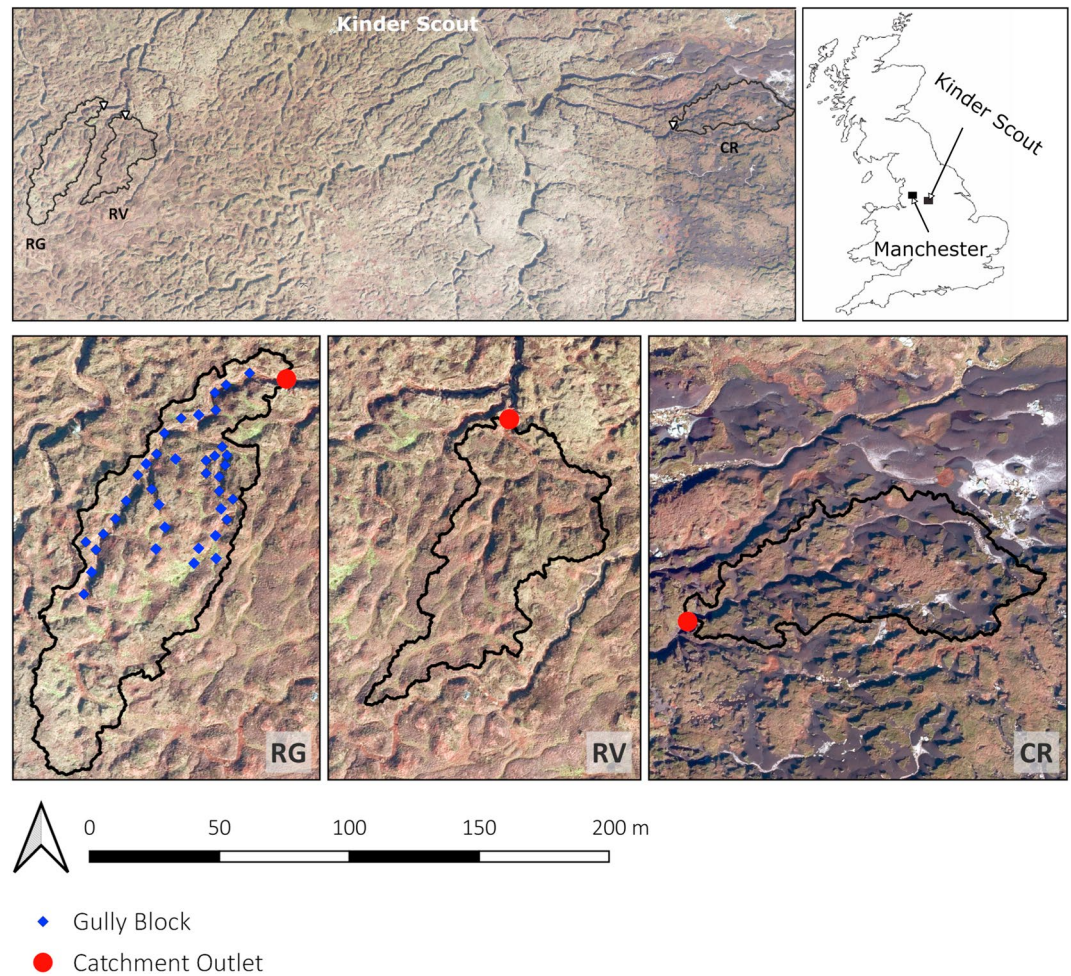
Figure 1. Study catchment outlines overlain on satellite images (obtained in 2020) to show: bare peat cover at the control site (CR), extensive revegetation at the intervention sites and the locations of gully blocks at the revegetated and blocked site (RG). Dark brown areas are bare peat; white areas are mineral bedrock exposed by total removal of the blanket peat.

2. Study Sites

The BACI experiment of Shuttleworth, Evans, Pilkington, et al. (2019) used rainfall and discharge data over a 5-year period for three micro-catchments located on the Kinder Scout Plateau in the Southern Pennines, UK (see Figure 1).

Kinder Scout is one of the most severely eroded areas of blanket peat in the UK, characterized (before intervention) by extensive areas of bare peat flats and networks of erosion gullies (Figure 1). The three micro-catchments were selected to be as close to each other as possible and have comparable geometries, erosion and gully characteristics prior to interventions (Table 1). Intensive monitoring began in June 2010 with a preintervention period of 15 months, followed immediately by more than 32-months of post-intervention monitoring. One catchment was retained as a control (CR) while interventions were implemented in the other two: revegetation alone (RV) and revegetation with gully blocking (RG). Between 2010 and 2014, revegetation led to 76% reduction in bare peat cover, with no significant change at the control site.

Table 1
Before-After-Control-Intervention Catchment Characteristics (Shuttleworth, Evans, Pilkington, et al., 2019)

Attribute	Unit	CR	RV	RG
Total area	m^2	7,008	4,468	7,096
Gullied area	%	32.9	22.9	28.5
Bare peat in non-gullied area	%	55	52	48
Max elevation	m	618	617	619
Min elevation	m	612	611	611
Mean slope	degree	10.1	8.9	9.7

Note. Values for gullied area and bare peat in non-gullied area, are in percentages of the total area.

Gully blocking at RG involved: 17 0.5 m high stone dams composed of loose piles of cobbles (75–200 mm diameter), across the width of the gully, spaced c. 6–7 m apart; and 20 timber dams in smaller tributary gullies of similar height but with a 38 mm deep "V-notch" cut into the top board. The resultant block density (50 blocks per hectare), configuration and construction are all typical for restoration activities in this area. Note that timber dams generally retain a pond with its surface at, or close to the base of their V-notch. At the time of installation stone dams were permeable, but became progressively clogged with eroded peat deposits. Thus, the permeability of the cobble stone dams over our study period is not clear, but we will discuss the implications of this uncertainty in the discussion section.

We chose a 6-week study period starting from August 22, 2010 as the before intervention period, and the same period in 2012 as the after intervention period. These periods hereafter are referred to as "before" and "after" periods. Choosing the same period for each year reduces the possibility that seasonal variations play a role in the hydrograph behavior. Other factors that informed our choice of study periods are: (1) large storms, with return periods within the largest 10% in the past 10 years, were present in both periods; (2) temperatures were always above zero; and (3) runoff coefficients for storms in the study period were never greater than one (a hydrograph mass balance check to identify gross errors in rainfall or discharge measurement).

To put our storms into context we both: (1) compare storm rainfall intensity and duration at our sites with that associated with nearby flooding; and (2) compare flood peaks resulting from our storms on nearby larger rivers to those at which flooding occurs. The three discharge gauges that we used were: Compstall (catchment area 156 km^2), Yorkshire Bridge (126 km^2), and Uppermill (41 km^2). We defined a lower bound on the rainfall intensity and duration at which flooding occurs based on reported rainfall for the three events in our record. Namely, storms in panels (a and b) in Figure 5, from our 2010 period; and storm depicted in panel (c) from our 2012 record. This analysis suggests that: (1) none of the storms within our study period resulted in flooding downstream of our micro-catchments; (2) the two 2010 storms together had an average intensity of 0.7 mm/h over 72 h (50% of that associated with flooding); (3) the 2012 storm (in Figure 5c), had an intensity and duration (1.4 mm/h over 54 h) that has previously led to flooding in the region; (4) the largest 2010 storms (Figure 5a), though less extreme in intensity-duration terms, still resulted in a downstream flood peak 79% of that responsible for flooding at the Uppermill gauge; (5) the 2012 storm resulted in flood peaks 62%, 77%, and 75% of that responsible for flooding at Compstall, Yorkshire Brige, and Uppermill, respectively. We discuss the implications of this for our findings in Section 7.

3. Numerical Model

TOPMODEL (Beven & Kirkby, 1979) and its variants have been widely applied in rainfall-runoff modeling of peat catchments in many parts of the world (e.g., Gao et al., 2015, 2016, 2017; Lane & Milledge, 2013; Stocker et al., 2014). The full set of principles and mathematics underlying TOPMODEL have been discussed in detail by many authors including Beven and Kirkby (1979), Keeland et al. (1997), and Kirkby (1997), and will not be explained here. However, the core assumptions of the model are (Kirkby, 1997): (1) phreatic surface is always parallel to the surface, thus the hydraulic gradient in the saturated zone can be approximated using the local topographic slope ($\tan\beta$); (2) transmissivity decays exponentially with water table depth (or local storage deficit); and (3) subsurface flow can be modeled as a series of steady-states. TOPMODEL's assumptions fit well to the case of blanket peat catchments of interest, where runoff is dominated by surface or near surface flow and the rate of flow declines rapidly with depth in the top few centimeters of soil profile (Holden & Burt, 2002).

In its classic form, TOPMODEL does not distinguish between the connected and disconnected saturated areas, meaning that whether or not a saturated patch is connected to the drainage network, its overland flow is routed to the outlet. Lane et al. (2004) showed that this can lead to over-estimation of runoff generation and proposed a Network Index version of TOPMODEL that only routes the connected portions of saturated areas as overland flow. Later, Lane and Milledge (2013) coupled this Network Index version of TOPMODEL with spatially distributed unit hydrograph (SDUH) routing to account for the timing of delivery of water from different parts of the catchment to the outlet. They showed that even this simple spatially and temporally averaged treatment of overland flow was capable of considerably improving model performance

for upland peat catchments. Therefore in this study we use the Lane and Milledge (2013) model, which improves on the original TOPMODEL while keeping its major advantage of being very CPU-efficient for calibration purposes.

This version of TOPMODEL has nine uncertain parameters (Table 2). We set the parameter ranges based on a combination of literature review and previous experience. We chose ranges that were as wide as possible while remaining physically meaningful. We have chosen $S_{rz_{max}}$ and t_d to be comparable to those previously used by Lane and Milledge (2013) for a UK blanket peat catchment. In our experience m and T_0 are the most sensitive parameters (as also found by Kavetski et al. [2003]) and are catchment-rainfall dependent, therefore it is generally not possible to use values calibrated to other catchments. Thus, we performed a manual sensitivity test to obtain the most conservative ranges beyond which the model is insensitive to m and T_0, when input individually or together.

Overland flow celerity (c) represents the speed with which rainfall-induced perturbations in the surface water travel overland and downstream. It is related to, but different from velocity which is the speed of travel of water particles as measured by tracers. Although both terms have often been used interchangeably, Beven (2020) demonstrates their intrinsic differences and shows that celerity is the relevant property when hydrograph response is of interest (e.g., rainfall-runoff studies); he also shows that celerity in downstream direction is always ≥velocity. We will treat c as a spatially and temporally averaged celerity over the entire catchment area.

Note that the study catchments are small, such that the longest travel distance for water flowing overland to the outlet in all three catchments is 293 m. The observational rainfall-discharge time series was recorded at 10-min intervals, which dictates the timestep in the model. Thus, the combination of a maximum travel distance and a fixed timestep, sets an upper bound on c of 0.5 m/s. Average celerities exceeding 0.5 m/s result in all overland flow response reaching the outlet within one time step, thus the model becomes insensitive to values of c greater than 0.5 m/s. We do not expected this to pose any limitation on our study because: (1) surface flows will be shallow for the majority the flowpaths; (2) measured velocities for such flows in peat catchments are less than 0.1 m/s (Holden, Kirkby, et al., 2008); and (3) a 1-D diffusion (e.g., Manning) based estimate suggests that celerity is less than twice the velocity (McDonnell & Beven, 2014).

Potential evapotranspiration is defined as the maximum possible evaporation and transpiration when the catchment is fully saturated and root zone storage is full. We follow Metcalfe, Beven, and Freer (2015) and account for annual and diurnal variation by assuming a linear relationship between insolation and potential evapotranspiration, and sinusoidal variation in insolation both over a year and between sunrise and sunset. We take this approach given limited available data because it captures expected temporal dynamics in potential evapotranspiration at a cost of only one additional parameter (E_p, representing the annual average potential evapotranspiration rate in units of meters per day) and has been found to provide comparable performance to more complex formulations for a number of UK sites (Calder et al., 1983).

We found that modeled discharge became insensitive to number of the HRUs (or topographic index classes), n_c, >20, therefore this value was used in all simulations. Further, we eliminated the initial subsurface flux (Q_0) and the initial root zone storage (S_{rz_0}) parameters by using a "spin-up" period, where the model was run using measured rainfall data for a period of 7-days prior to the study period (i.e., August 22, 2010/2012). The spin-up duration was chosen to include a large storm so that the subsurface water conditions, in particular the initial root zone store and the initial subsurface flux, were balanced. The six remaining uncertain parameters were: exponential transmissivity decay parameter m, transmissivity at saturation T_0, overland flow celerity c, annual average potential evaporation rate E_p, unsaturated zone time delay, t_d, and maximum root zone storage, $S_{rz_{max}}$.

4. Framework

4.1. Model Calibration Using GLUE

Hydrological models in general, and TOPMODEL in particular, are well known to experience "equifinality," where many different parameter-sets result in comparably good fits to observational data. Thus, we fo-

Table 2
Uncertain Model Parameters and Their Range

Parameter	Unit	Range
c: Overland flow celerity	m/s	$0.001 : 0.5$
E_p: Potential evapotranspiration rate	m/day	$10^{-6} : 0.05$
$S_{rz_{max}}$: Maximum root zone storage	m	$0.005 : 0.02$
t_d: Unsaturated zone time delay	hr/m	$1 : 300$
m: Exponential decay parameter	m	$5 \times 10^{-5} : 0.01$
T_0: Maximum transmissivity	m²/s	$10^{-9} : 2 \times 10^{-5}$

cus on parameter distributions and their changes following revegetation and gully blocking rather than seeking a single optimum parameter-set that characterizes each catchment under study. We apply a Generalized Likelihood Uncertainty Estimation (GLUE) methodology (Beven & Binley, 1992), using uniform prior distributions for each uncertain parameter, and running 160,000 TOPMODEL simulations with parameter-sets sampled randomly from these distributions. We retain only "behavioral" parameter-sets, for which modeled discharge is feasibly within error (limits of acceptability) of the observed discharge for 99% of the observation period.

Since the observed discharge record remains our best estimate of the true discharge (despite its associated uncertainties) we then weight "behavioral" sets based on their agreement with the observed record using three metrics: Nash-Sutcliffe efficiency (NSE), Root Mean Squared Error in magnitude of the 10 most distinct peaks (RMQ) and Root Mean Squared Error in timing for the same peaks (RMT). By examining and comparing the "weighted" distributions of each parameter in the "before" and "after" periods, we aim to identify the effects of interventions, and use this information to infer catchment functions through numerical experimentation. To calculate the likelihood weights assigned to each behavioral parameter-set, we evaluate a linear weight for each metric and multiply them to obtain an overall weight. Thus, for each metric M, the normalized weight is given by:

$$w_i^M = \frac{|M_i - M_0|}{\sum_{i=1}^{N_b} |M_i - M_0|} \quad ; \quad i = 1 : N_b \quad ; \quad M \in \{NSE, RMT, RMQ\} \tag{1}$$

where N_b is the number of behavioral parameter-sets and M_0 is the value of the metric for the worst performing set. The worst performing set is given the lowest weight and the best set the highest weight, while sum of all weights equal to one. The overall likelihood weight, which takes into account the likelihoods from each metric, is calculated by multiplying weights for all metrics and normalizing by the sum of values:

$$W_i = \frac{\prod w_i^M}{\sum_{i=1}^{N_b} \left(\prod w_i^M \right)} \quad ; \quad M \in \{NSE, RMT, RMQ\} \tag{2}$$

To establish how large a sample size should be to ensure sufficiently dense sampling of the six-dimensional parameter space, we: (1) generate three independent random-sets (S1, S2, and S3) each containing 5,000 randomly sampled parameter-sets; (2) run the model for each of S1, S2, and S3 and in each case calculate the weighted mean of each parameter (Equation 2); and (3) iteratively double the sampling size in each random-set until the weighted means of the parameter values in all three random-sets converge. Convergence occurs at or before a sample size of 160,000 for all parameters, study periods, and micro-catchments (CR, shown in Figure 2). In addition, weighted mean parameter values remain stable even as the sample size is doubled from 80,000 to 160,000. Both observations suggest that 160,000 samples provide sufficient sampling in our case.

We also performed a blind validation test, where the calibrated parameter-sets were used to predict runoff responses for a period (October 5–20, 2010 and 2012) adjacent but not overlapping the calibration period (August 22–October 4, 2010 and 2012). We found that none of the distributions of the objective functions were significantly different with 95% confidence, providing more confidence that our behavioral sets are representative of the sites being studied.

4.2. Obtaining Limits of Acceptability

Limits of Acceptability (LOA) are upper and lower bound limits on the observational discharge within which the model predictions must lie to be deemed "behavioral" and thus accepted. We seek to define LOA
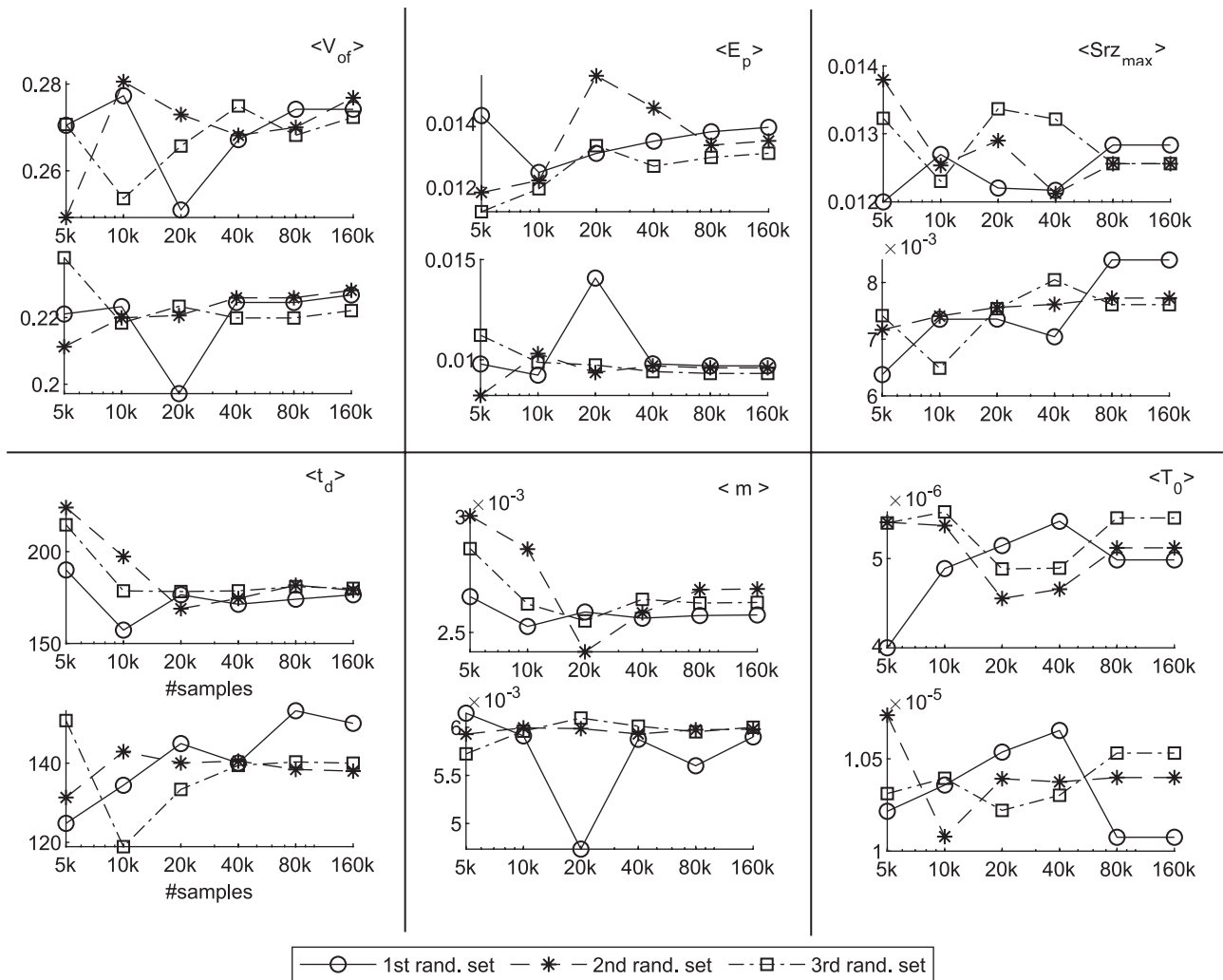
Figure 2. Results of the sampling convergence test for GLUE uncertainty estimation. In each panel top plot is the "before" and bottom plot is the "after" period.

to minimize Type II errors, where a good model is rejected as a result of errors in validation or input data (Renard et al., 2010). These errors can arise both from measurement limitations/errors, and from commensurability and scaling issues where the time/space scale of measurements differ from those within the model. We estimate errors in: water level measurement, rating curve, and in the rainfall measurements that drive the model.

Discharge from each catchment was estimated at 10-min resolution using a 90° triangular V-notch weir with stage behind the weir measured using pressure transducers. Stage was converged to discharge using a standard rating curve relationship of the form $\hat{q} = a\hat{h}^b$, where $\hat{q}$ is the observed discharge, $\hat{h}$ is the observed stage, and a and b are empirical parameters related to geometry and discharge coefficient. Following Keeland et al. (1997), and making the conservative assumption that measurement accuracy was relatively low for such equipment, we use a stage measurement error standard deviation of $\sigma = \pm0.0075$ m. We then propagate this stage error through the rating curve using a standard frequentist inferential methodology (Petersen et al., 2005) to calculate an uncertainty ratio R, which is the span of the 95% confidence limit in discharge, divided by the observed discharge:

$$R = \frac{\hat{q}_{95}}{\hat{q}} = \frac{3.92b\sigma}{\hat{h}} \tag{3}$$

where σ is the measurement error standard deviation. Therefore, the contribution of stage measurement errors in the LOA, calculated as upper and lower bounds on the discharge is taken to be $(1 + R)\hat{q}$ and $(1 - R)\hat{q}$, respectively.

Stage was converted to discharge in the BACI experiment using theoretical rather than empirical rating curve relationships, thus we propagate parameter uncertainty in theoretical rating curves through the weir equation of Shen (1981) relating observed discharge, $\hat{q}[LT^{-1}]$ and stage, $\hat{h}[L]$:

$$\hat{q} = C_e \frac{8}{15} \sqrt{2g} \tan\left(\frac{\theta}{2}\right) \hat{h}^{\frac{5}{2}} \tag{4}$$

where $C_e[-]$ is the non-dimensional coefficient of discharge, $g[L^2T^{-1}]$ is the gravitational acceleration, θ is the V-notch angle. For 90° weirs, the $C_e[-]$ value varies between 0.5 and 0.7 (Shen, 1981), therefore these values were used to obtain an upper and lower bound on discharge as maximum and minimum possible variation in discharge as a result of rating curve uncertainty.

Since observed discharge is estimated from an instantaneous stage measurement every 10 min but modeled discharge is the average of the volumetric water flux from the catchment over the 10-min timestep, there is a potential time-commensurability error between the two quantities. This error will be most severe under rapidly changing discharge but it is difficult to conceive of scenarios in which instantaneous discharge could differ from the 10-min average by more than it differs from the instantaneous measurement 10 min before or after. Thus the upper bound on discharge for timestep t is set to the maximum of the upper bounds in timesteps $t - 1{:}t + 1$, unless t is a trough, when the upper bound is set to the minimum of $t - 1$ or $t + 1$. The lower bound for timestep t is set to the minimum in timesteps $t - 1{:}t + 1$, unless t is a peak, when the lower bound is set to the maximum of $t-1$ or $t + 1$.

There is no commonly accepted method of evaluating rain gauge error, and particularly how it manifests itself in the outlet discharge predictions when the erroneous rain gauge data is input to a numerical model to predict a discharge (Vrugt & Beven, 2018). Rainfall measurement errors stem from errors in the measurements themselves and from their inability to capture spatial and temporal variability (McMillan et al., 2012). Since each catchment contains only a single rain gauge recording at only 10-min frequency we cannot estimate the spatial and temporal variability of rainfall in our study catchments. However, the catchments are sufficiently small (order of 10^{-3} km^2) relative to storm-cloud footprints (Féral et al., 2006; Sauvageot et al., 1999) and the recording frequency sufficiently high relative to storm durations, that we expect errors due to spatial and temporal variability to be small. Finally, point measurement errors can originate from different sources such as the time-sampling effect caused by the discrete character of the tipping bucket measurements, water flow instabilities in the gauge funnels, and turbulent airflow around the rain gauges caused by wind (Pollock et al., 2018; Rodda, 1967). The BACI experiment has provided three rain gauge recordings, one for each catchment, which are within 400 meters of each other ($P_{CR}(t)$, $P_{RV}(t)$ and $P_{RG}(t)$). We use these to estimate rain gauge measurement errors by following Ciach (2003) in assuming that the average of rain gauge measurements is likely to be closer to the true value than any of the individual recordings, due to smoothing of local random errors. Thus, if $\overline{P}(t)$ is the mean value of the precipitations, then, for the given timestep, the error is taken to be the maximum deviation from the mean:

$$e_P(t) = \max\left(|\ \overline{P}(t) - P_i(t)\ |\right) \qquad i \in \{CR, RV, RG\} \tag{5}$$

To include the contribution of rainfall measurement errors in the limits of acceptability, they must be translated into discharge error. However, non-linearity in rainfall-runoff models leads to accumulation of errors in state variables, resulting in complex and non-traditional time series of residuals (Vrugt & Beven, 2018), and making simple error propagation unfeasible. Instead, we choose to propagate the $\overline{P}(t) \pm e_P(t)$ through the numerical model and assess its magnitude of impact on discharge. Since, this must be performed prior to model calibration, we run 160,000 simulations by sampling parameters from the prior distributions. We repeat this exercise for the same parameter-sets using three different rainfall inputs: (a) $\overline{P} - e_P$, (b) $\overline{P}$ and (c) $\overline{P} + e_P$. The resulting discharge prediction values for each rainfall case are then averaged across all the

parameter values; the distance between averaged discharge values of (a and c) compared to (b) is taken to be the upper and lower bound as a consequence of rainfall measurement errors.

Thus, the limits of acceptability were calculated by adding the effects of all of the above mentioned errors, in the form of upper and lower bounds on the observational discharge.

Note that there are other errors (sensor drift and V-notch blockage) that we have not been able to quantify and include in LOA. Field observations suggest that sensor drift was negligible and that blockages are both rare and short-lived, but can considerably distort discharge predictions. We have sought to account for these by retaining models even when their predictions fall outside LOA for up to 1% of the timesteps.

4.3. Obtaining Parameter Distributions

Figure 3 shows the step by step process that was undertaken in obtaining the parameter distributions which characterize each catchment-period under study. For the purpose of illustration, the potential evaporation parameter, E_p, for the CR catchment and in the "before" period is shown here, but the procedure is the same for all parameters, catchments and periods.

A uniform random distribution of 160k samples is input to the model (Figure 3a), for each of which the model generates a discharge prediction. The parameter values associated with model predictions that sit outside of the LOA are discarded (Figure 3b). Note that since discharge predictions that remain within the LOA throughout the entire study period are extremely rare (Beven, 2006; Freer et al., 2003) we have allowed for 1% of predictions to be outside the LOA (similar to Blazkova & Beven, 2009; Van Straten & Keesman, 1991). In Figure 3c, an extra condition is imposed on model predictions to reflect field observations in UK upland blanket peatlands that discharge response to rainfall is dominated by overland flow (Daniels et al., 2008; Holden, 2009; Holden & Burt, 2002, 2003). Thus the overland flow (OF) criteria requires that >70% of predicted outlet discharge should be composed of overland flow (a conservative, i.e., low, value based on field observations). Finally, after the model predictions are filtered by the LOA and OF criteria, the remaining (behavioral) parameter-sets are ranked according to model performance and linearly weighted (Equation 2) such that better performers get higher weights. Using these weights, the weighted counts of the histogram bins are calculated, which are then normalized by dividing by the largest value of the weighted counts to obtain the Normalized Weighted Counts (NWC), which vary between 0 and 1. These normalized weighted distributions (Figure 3d) enable comparison between catchments' distributions. All subsequent parameter distributions are normalized weighted distributions, their *y*-axis range is always 0 and 1, and thus is not shown.

5. Results

5.1. Parameter Distributions Before and After Interventions

Using the procedure described in Section 4.3, the normalized weighted distributions for all parameter-catchment-period combinations are generated and shown in Figure 4. The top and bottom rows of each panel contain the "before" and "after" distributions respectively. Note that even at the control site, where there has been no intervention, the parameter distributions differ between the "before" and "after" periods. Shuttleworth, Evans, Pilkington, et al. (2019) also observed changes in hydrological behavior at the control site (emphasizing the need for a BACI experiment) and referred to it as "natural temporal variability." Thus distributions should be interpreted relative to the control site in each case.

These distributions highlight the parts of the parameter space that performed better while satisfying the LOA and OF criteria. In GLUE terms, these distributions can be interpreted as indicating the likelihood that particular parts of the parameter space represent the true values for the catchment of interest.

Examining the shifts in parameter distributions following interventions and starting from the top left panel of Figure 4, distributions of overland flow response (celerity, *c*) in the "before" period differ slightly among catchments, however, all three assign almost zero likelihood to *c* values smaller than 0.1 m/s. There is an
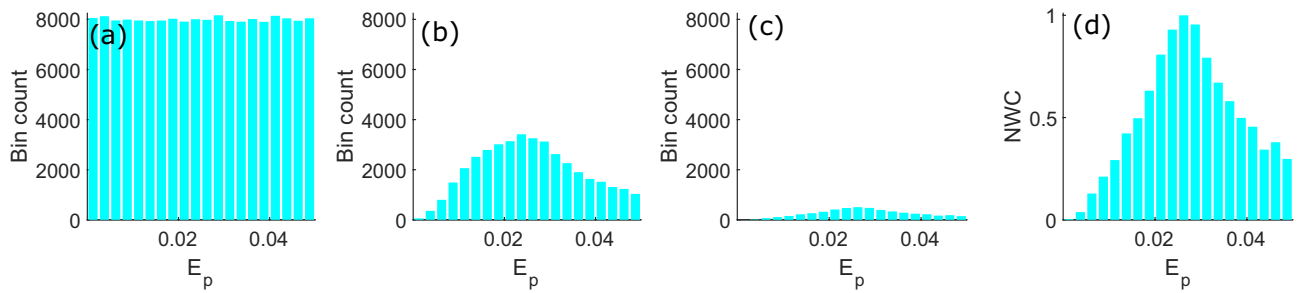
Figure 3. Step by step demonstration of how parameter distributions were obtained. Example for evapotranspiration parameter, E_p. NWC stands for Normalized Weighted Count of the histogram bins.

overall reduction in surface flow celerity in the "after" period in all sites, but the reduction is larger in RV and RG than in CR, and is largest in RG.

Root zone storage $S[L]$, in the model, accounts for overall interception storage of vegetation, within the root zone or in surface ponds (whether natural or behind gully blocks). It has a maximum capacity $S_{rz_{max}}$; is emp-
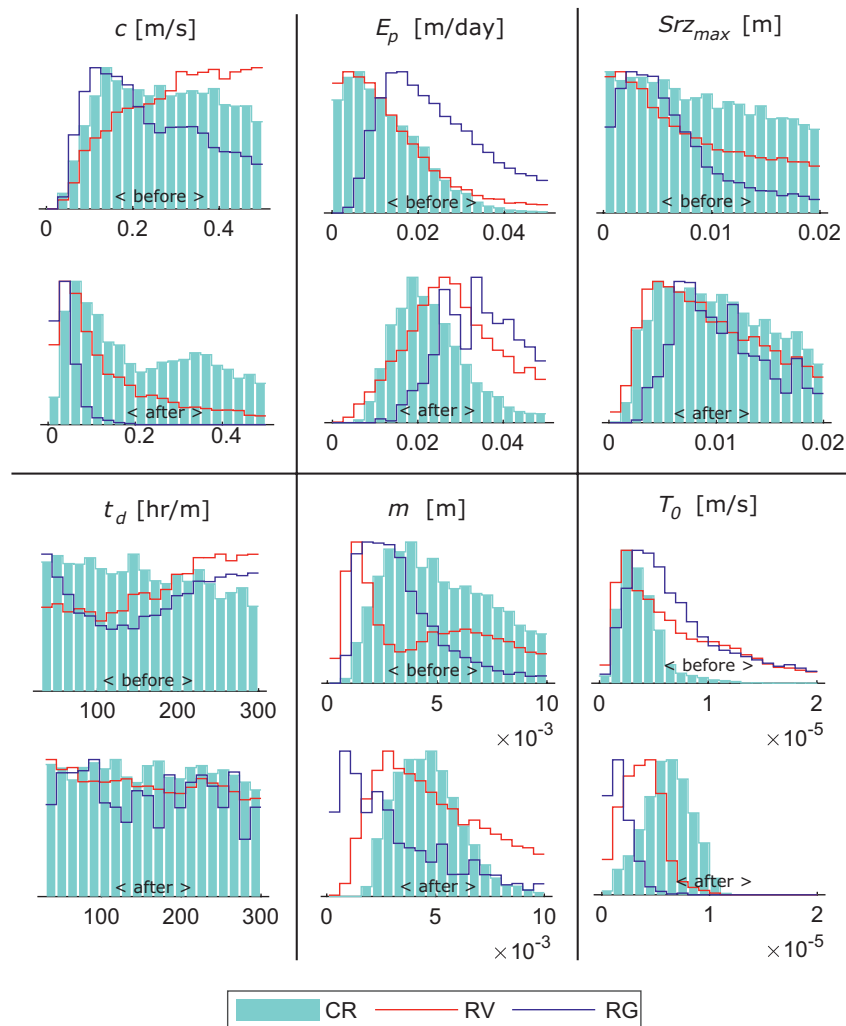

Figure 4. Before-After intervention parameter distributions for all three catchments under study. *CR* denotes the control, *RV* denotes the revegetated and *RG* denotes the revegetated and gully blocked case. Top row are the "before" distributions and bottom are the "after". The *y*-axis range is 0 and 1 for all distributions and thus is not shown. For description of parameters see Section 3.
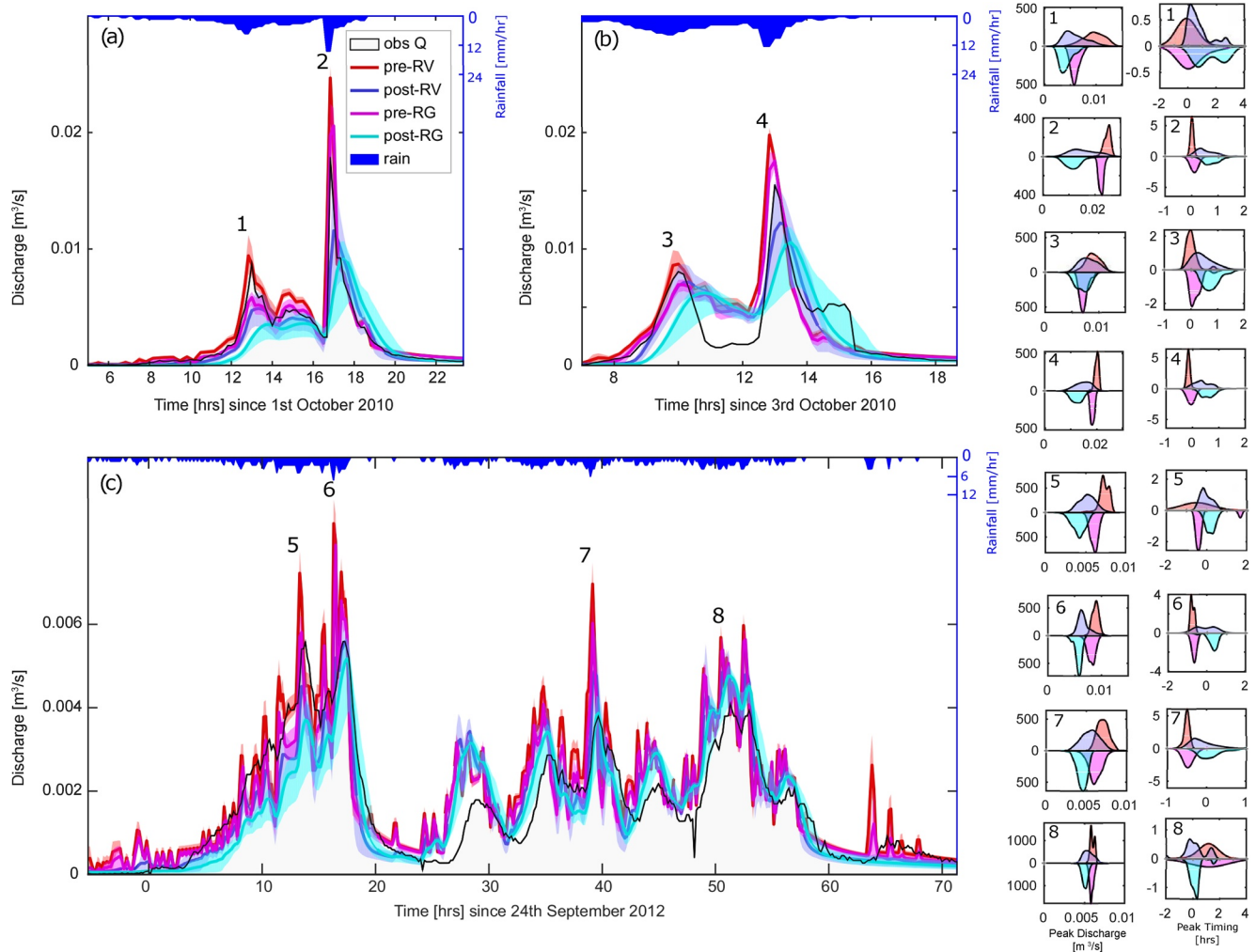
Figure 5. Absolute impact experiment: predicted discharge under four different model parameterizations (panels a–c). Numbers 1–8 are peak ID numbers. Right of the figure are probability density functions (PDFs) of each peak.

tied only by evapotranspiration at a rate $E_a = E_p \dfrac{S}{S_{rz_{max}}}$; and is the only source of evapotranspirative loss in

the model. Unless this store is full rainfall cannot contribute to either vertical infiltration or surface flow. Before the interventions, the distributions of potential evapotranspiration (E_p) are very similar in CR and RV, while RG is characterized by higher E_p. After the interventions, all sites shift toward higher E_p values, with RG larger than CR and RV slightly larger than RG. Before the interventions, the distributions of $S_{rz_{max}}$ are very similar in all sites. After interventions, all sites shift toward higher $S_{rz_{max}}$, the increase is similar for CR and RV, while RG shows a clear increase relative to CR.

The unsaturated zone time delay parameter (t_d) controls the vertical flux of water from the peat surface to the saturated zone below. The model is able to produce good fit to the observed discharge across the full range of unsaturated zone time delays both before and after interventions and at all sites. This indicates that the model is not sensitive to this parameter. This is perhaps due to the very shallow water table depth in peat catchments and their propensity to form saturated areas during a storm, limiting both the vertical and lateral extent of the unsaturated zone.

The parameters m and T_0 control the catchment's transmissivity-depth profile and thus the relationship between subsurface flow and saturation deficit. T_0 sets the maximum transmissivity when the peat is fully

saturated, while m controls the rate of exponential decay in transmissivity as saturation deficit increases (higher values indicate slower decay). Before the interventions, T_0 values are similar at all sites. After the interventions, T_0 shifts toward higher values for RV, but lower values for RG, though this difference may be offset by an increase in m for RV and decrease for RG after interventions.

Care is required in interpreting these parameter distributions. They show the magnitude and direction of individual parameter shifts in response to calibration to each catchment-period pair, but it is difficult to infer how significant these changes are when parameter "sets" are used to predict runoff. This is because the impact of a parameter shift (e.g., in E_p) is controlled by both its magnitude and the model's sensitivity to that parameter. Since parameters interact this will also depend on the values of other parameters. This is clearly true for m and T_0, which together control transmissivity but is also important for evapotranspiration, which is controlled by both E_p and $S_{rz_{max}}$. For this reason, in the following sections we perform numerical experiments to investigate how the model's runoff predictions change in response to varying sets of "parameter-sets," each of which pertains to a catchment-period pair.

5.2. Numerical Experiment #1: Absolute Impact

The first numerical experiment uses the behavioral parameter-sets of Section 5.1 to examine the hydrograph response of each catchment in terms of peak magnitudes and timings. In order to isolate the impact of the interventions we create catchments that are "virtual twins." These twins have the same topography, flow pathways, input rainfall, and antecedent conditions and differ only in the set of behavioral parameter-sets that represent each intervention.

First, we used RG's topography, and input rainfall for the 2010 study period to drive the model. The model was then evaluated with four different sets of "behavioral" parameter-sets: (1) pre-RV and (2) post-RV representing the pre- and post-intervention conditions at the revegetated site; (3) pre-RG and (4) post-RG representing pre- and post-intervention conditions at the revegetated and gully blocked site. We then repeated this process but using the RG's 2012 rainfall. Finally, we repeated the whole process but using RV's topography, and RV's rainfall from the 2010 and 2012 study periods (i.e., all combinations of the two rainfalls and two topographies).

The results are shown in Figure 5. Throughout this figure cooler colors represent preintervention conditions and hotter colors represent post-intervention. Due to the length of the records we only show the largest storms in the record (belonging to the 2010 "before" period) in panels (a and b), and the most complex storm (belonging to the 2012 "after" period) in panel (c). The middle line of each color band is the mean of all predictions and the width of the bands are indicative of the variation in the predictions across all behavioral parameter-sets. The numbers 1–8 are peak ID numbers and the smaller panels on the right of the figure are the associated probability density function (PDF) of each peak discharge and peak timing prediction (left and right columns, respectively). In each of the small boxes, upper PDFs are for the RV case and lower (upside down) PDFs are for RG. These PDFs show how (in)significant the effects of interventions have been on each peak, such that the more overlap there is between the pre- and post-intervention PDFs, the less significant the effects have been for that particular peak, and vice versa.

Post-intervention hydrographs are noticeably smoother than preintervention (Figure 5). This reduced amplitude and increased wavelength of the discharge timeseries is likely due to reduction in the speed of surface flow response (celerity) associated with increased surface roughness following revegetation and gully blocking. Interventions almost always reduce peak discharge and delay the timing of the peaks. This attenuation is present in both large and small peaks (e.g., Peaks 1 and 2) though it appears less marked in some peaks (e.g., Peak 3) than others. Reduced attenuation of Peak 3 is also clearly illustrated by the lack of separation between its pre- and post-intervention PDFs, particularly for peak magnitude and partially for timing (Figure 5).

To investigate whether there is a systematic relationship between storm size and peak discharge/lag time change, we examined the 40 most prominent peaks in the record (10 from each period-topography combination, Figures 6a and 6b). For each peak, we used the weights from Equation 2 to calculate the weighted-aver-

age of the behavioral model predictions pre- and post-intervention (Q_{pre} and Q_{post}, respectively). The post-intervention change in peak discharge is then calculated as a percentage of the Q_{pre} using: % = (Q_{post} − Q_{pre})/ Q_{pre} × 100, which is the *y*-axis of Figure 6a. For peak timings, the same procedure is followed but without normalization. Thus absolute changes in timing are shown on the *y*-axis of Figure 6b.

Note that in obtaining Figure 6 sometimes precise identification of the timing of the hydrograph peaks was not possible. Delays and attenuations in the predicted hydrographs cause some peaks to interfere or merge making it difficult or impossible to identify corresponding peaks in "pre" and "post" model evaluations. This difficulty is demonstrated by the peak timing PDF of Peak 8 (Figure 5), where PDFs of RG show considerable "decrease" in lag time following revegetation and gully blocking which is inconsistent with the hydrographs for Peak 8 (Figure 5c). Peak magnitude change is not affected by this issue, because in this case it is only the magnitude of the peaks that matter and thus peaks do not necessarily need to correspond exactly. However, this introduces considerable uncertainty in estimates of timing change. To minimize these uncertainties we manually identify "complex" multi-modal peaks (such as Peak 8), remove them from peak timing analysis (Figure 6b) and flag them in our peak magnitude analysis (as circles rather than squares in Figure 6a).

Both interventions resulted in significant (95% CI) reduction in peak discharge for all storms independent of storm size (Figure 6a). Even Peaks 3 and 8, which show least reduction are significant at 95% confidence. Mean discharge reduction was 48% for RV and 58% for RG suggesting that gully blocking results in further reductions in peak discharge (above those from revegetation alone). These mean values aggregate considerable inter-storm variability with 90% reduction in peak discharge possible in some of the smaller storms but less than 20% reduction in four storms for RV and two for RG. The 2010 study period includes the two largest storms (Peaks 2 and 4 in Figure 5). These storms have rainfall intensity-duration properties that suggest an approximately annual return period (Jones et al., 2010) and peak discharges within the largest 10% over 10 years of observations at the control site. These two storms have modeled discharge reductions of 30%–40% for RV and 40%–55% for RG.

For the non-complex storms for which timing analysis was possible, interventions always increased lag times (Figure 6b). The lag time increase was always >10 min for revegetation alone and >30 min for revegetation and gully blocking, with overall mean (across 40 peaks in 2 periods for 2 topographies) of 0.47 h (28 min or 148%) in RV and 0.83 h (50 min or 265%) in RG. For the two largest storms, the lag increase in RV was roughly 0.5 h (30 min or 159%) and in RG was around 0.9 h (55 min or 280%).

Note that we also performed the above experiment for the control site (CR), however, the pre-CR and post-CR parameter-sets resulted in discharge predictions well within the pre intervention range (i.e., no significant lag increase or peak reduction), providing more confidence that our model is able to capture the observed behavior of these catchments. Given the near identical discharge predictions of CR in both periods, comparing the post-RV and post-RG results to pre-RV and pre-RG, respectively, would be the same as comparing them relative to CR; meaning that our results remain comparable to those experimentally obtained by Shuttleworth, Evans, Hutchinson, and Rothwell (2015), who calculated changes relative to CR. A version of 5 with both sets of CR results is given in the electronic supplement to this study for reference.

5.3. Numerical Experiment #2: Relative Impact

The Absolute impact experiment estimated the impact of interventions when all parameters are changed from pre-to post-intervention values and Figure 4 indicates the magnitude and direction of parameter changes. However, the relative importance of each parameter in driving the intervention impacts remains unclear. The second numerical experiment aims to identify the parameters (and therefore processes) driving the observed effects of a particular intervention.

The experimental design is similar to the absolute impact experiment, holding rainfall and topography constant and varying only the input parameter-sets (the "behavioral" parameter-sets for each intervention). However, here we perform a one-at-a-time parameter switch, where a single parameter (or pair of parameters) is switched to its post-intervention values while all others are kept at their preintervention values. Changes in the predicted peak magnitude and timing can then be assigned to the switched parameter. We calculate the associated discharge change as a percentage of the post-intervention discharge where "all"
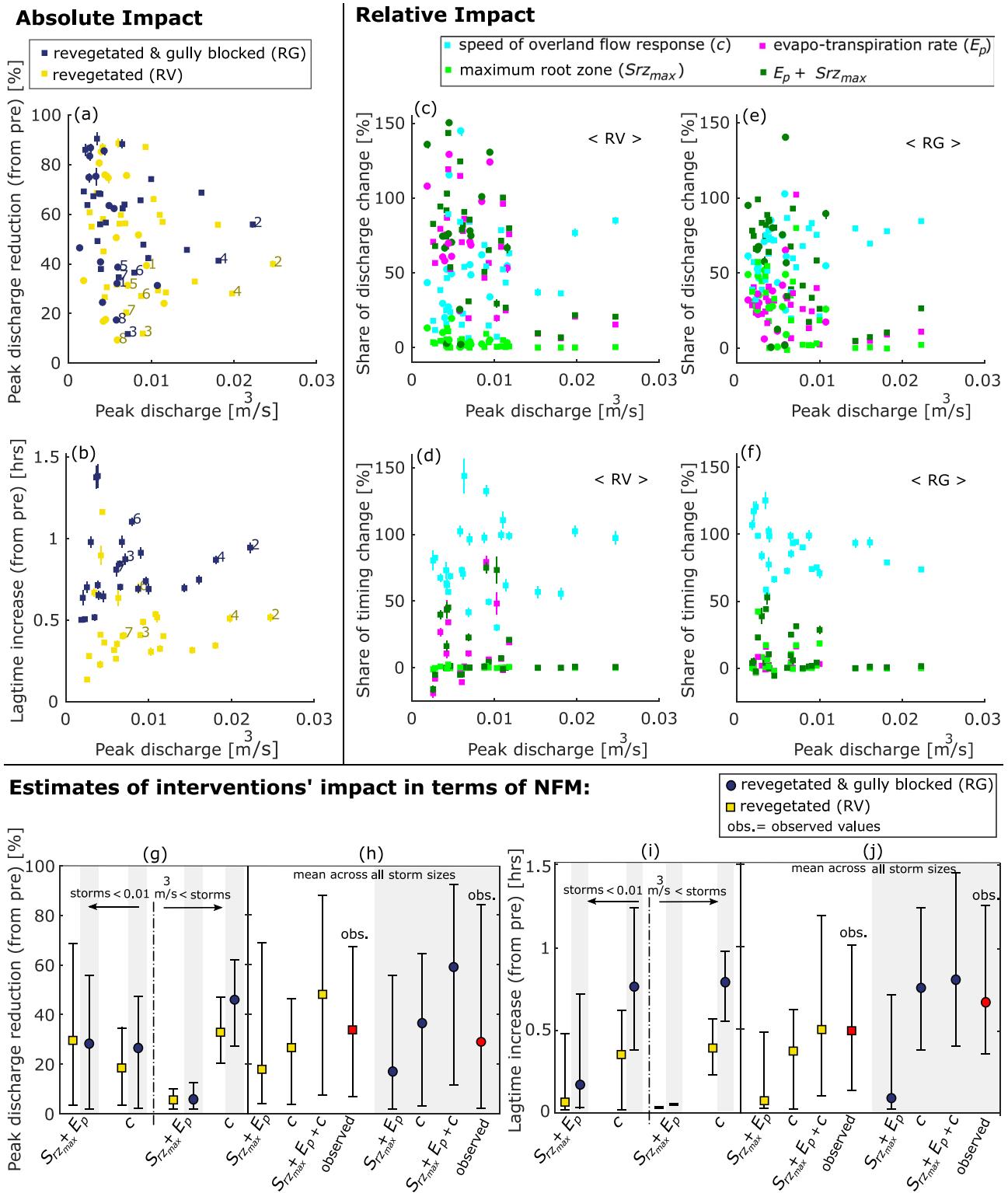
Figure 6. Absolute impact experiment: absolute magnitude of intervention impacts for different storm sizes and in terms of: (a) peak magnitude reduction relative to pre, (b) peak timing delay relative to pre. Relative Impact experiment: "relative" (to one another) impact of restoration intervention for different parameters and storm sizes: (c and d) parameter shares in peak reduction and lag increase, respectively, in RV. (e and f) for RG. NFM impact of processes: (g and i) magnitude reduction and lag increase, respectively, for different storm categories. (h and j) mean impact across all storm sizes and comparison with observed means.

parameters are switched to their "post" values. To do this we calculate the weighted-average (Equation 2) of all post-intervention discharge predictions for the case where only one parameter p is switched and all other are kept at "pre" values ($Q_{\text{post}.p}$). We then weighted-average the discharge predictions for the case where "all" parameters are switched to their "post" values ($Q_{\text{post}.all}$). The percentage change is then given by: $\% = (Q_{\text{post}.p} - Q_{\text{post}.all})/Q_{\text{post}.all} \times 100$ (Figures 6c and 6e). For peak timing a similar procedure is followed except that the results are not converted to percentage but shown as absolute differences (Figures 6d and 6f). Note that it is possible for the discharge/timing change due to a single parameter switch to be larger than that when all parameters are at the post-intervention values, because by switching parameters one-at-a-time we ignore parameter interactions, which might be counteracting one another. Thus the y-axis in Figures 6c–6f cannot be interpreted as the share of intervention impact in absolute terms, rather as a relative indicator of that parameter's importance.

Lateral hydraulic conductivity related parameters (m and T_0) when switched in together (because they jointly represent transmissivity) were found to have no significant share (with 95% CI) in delaying the flow or reducing the peak discharge, indicating that transmissivity is not altered two years after interventions; similarly for the vertical transmissivity parameter (t_d). Therefore, we assume the remaining parameters c, $S_{rz_{max}}$ and E_p are responsible for the total impact of each intervention and show only these parameters in Figures 6c–6f. Additionally, since $S_{rz_{max}}$ and E_p are closely linked through their influence on E_a (see Section 5.1), we study the overall effect of rootzone storage using the paired parameters $S_{rz_{max}}+E_p$.

In both treatments, and for the larger (>0.01 m³/s) peaks, discharge reduction is predominantly controlled by celerity, although the impact of $E_p+S_{rz_{max}}$ is still significant (>10% of celerity's impact). For smaller peaks (<=0.01 m³/s) discharge reduction is controlled primarily by $E_p+S_{rz_{max}}$. In terms of the relative role of E_p and $S_{rz_{max}}$, in RV, E_p dominates, while in RG, both are important. Unlike peak discharge plots, peak timings in panels (d and f) are more consistently controlled by celerity as the primary delaying mechanism independent of storm size or intervention. The effect of celerity is more pronounced in RG than RV (double on average).

5.4. Estimating the NFM-Potential of the Restoration Interventions

In their BACI analysis Shuttleworth, Evans, Pilkington, et al. (2019) noted significant variability in catchments' behavior. Two important sources of uncertainty are: (1) differences between catchments in topography and material properties of surface and subsurface flow pathways; and (2) differences between rainfall characteristics in the pre- and post-intervention period. By assessing deviations from the control site, and examining a large sample of storms Shuttleworth, Evans, Pilkington, et al. (2019) were able to detect the impact of restorations despite this variability. However, it is not clear how much of the observed effects were due to differences between catchments or in rainfall characteristics. Fixing topography and rainfall within the absolute impact experiment will reduce, but not eliminate these uncertainties because the parameter-sets were calibrated to the observed rainfall-discharge time series.

Unlike the absolute impact experiment, the magnitude of the changes predicted using the second experiment are not absolute. To estimate the absolute impact of each parameter (≈process) on peak discharge and lag times we combine the "absolute" impact magnitude of interventions predicted by experiment #1, with the "relative" impact magnitude of parameters predicted by experiment #2. Since m, T_0 and t_d had negligible effects, we exclude them from this analysis. Due to the coupled nature of E_p and $S_{rz_{max}}$ we treat them as a pair. Thus, we assume that the total absolute magnitude of interventions predicted by the experiment #2, is shared between the speed of overland flow response c, and the parameter pair $E_p+S_{rz_{max}}$.

To obtain an estimate for the absolute impact magnitude of each intervention that can be assigned to c and $E_p+S_{rz_{max}}$: (1) for each of the 40 peaks, and each parameter, we calculate a weight by dividing that parameter's share (from Relative Impact experiment) by sum of shares of both parameters, such that weights add to one; (2) for each peak and each parameter, we multiply this weight by the absolute impact magnitude of the intervention (from the absolute impact experiment); (3) for each parameter, we calculate mean values for peaks <0.01 m³/s, peaks >0.01 m³/s, and for the full sample of storm sizes. We then compare these to

the case where all three parameters c, E_p, and $S_{rz_{max}}$ are switched to their post-intervention values (a full parameter switch), and to the observed values (Figures 6g–6j).

In Figures 6g–6j, dots are the mean and the error bars represent the range (min–max). Panels (g and i) show that, celerity, c, alone is almost entirely responsible for lag time increase, regardless of storm size or intervention. Peak reduction is dominated by c in large storms regardless of intervention. In RV and in smaller storms $E_p + S_{rz_{max}}$ have slightly more impact in reducing the peak magnitude than does c. But in RG, and even for smaller storms, c reduction due to surface roughness is the dominant process reducing peak magnitudes.

Figures 6h and 6j show the mean of impact magnitudes across all storm sizes, for c and $E_p + S_{rz_{max}}$ individually, and for their combination. When comparing the impact of $c + E_p + S_{rz_{max}}$ to the observed mean from the BACI experiment, there is considerable overlap in the predicted and observed range of peak discharge and lag times. This agreement between the two alternative (mechanistic and statistical) analyses is encouraging. Note that, using only 2012 as the "after" intervention period (modeled here), the observed reduction in peak discharge is smaller at RG than RV. This is unexpected given the additional gully blocking undertaken in RG and is not consistent with other years (2013 and 2014, see Shuttleworth, Evans, Pilkington, et al., 2019) where the reduction in peak discharge was significantly larger at RG than RV. However, the mean of modeled peak magnitude reductions in RV is 10% and in RG is 30% more than the observed mean, which is more consistent both with our expectations (that gully blocking should provide additional NFM benefit) and with results from the succeeding years. This suggests a possible underestimation of the impacts by the BACI experiment due to experimental limitations/noise.

6. Discussion

6.1. Process Representation: Static Versus Kinematic Surface Storage

Shuttleworth, Evans, Pilkington, et al. (2019) showed that water tables at treatment sites rose following both revegetation and gully blocking, which resulted in a reduction in catchment subsurface storage, leading to an enhanced production of saturation-excess surface flow. Despite the increase in surface flow production, lag times increased and peak discharges decreased, suggesting that: (1) the additional surface flow is stored in some form of "surface" storage, (2) this increase in total surface storage more than offsets the effect of reduced subsurface storage due to water table rise. Here we distinguish between two different types of surface storage: static storage and kinematic storage.

6.1.1. Static Storage

Similar to Medici et al. (2008), Bussi et al. (2013), Peña et al. (2016), Ruiz-Pérez et al. (2016), and Buendia et al. (2016), we also define surface "static storage" as the portion of surface water that is intercepted (and immobilized) by vegetation and in-channel gully blocks, and is only emptied through evapotranspiration. During a rainfall event, static storage can intercept some of the rainfall and reduce hydrograph peaks. It may also delay the peaks by a lag proportional to the time it takes to fill the store. This storage is represented by the root zone storage variable, $S[L]$, in the model, with a maximum capacity $S_{rz_{max}}$ and emptying rate $E_a = E_p \dfrac{S}{S_{rz_{max}}}$. Together, the paired parameters $S_{rz_{max}} + E_p$ capture the overall behavior of the static storage unit; individually they provide useful information on the interplay of storage capacity and evapotranspiration in the static storage unit. Note that due to its immobility, static storage also represents the portion of rainfall that does not contribute to runoff. Therefore, increasing static storage capacity can reduce runoff coefficients (ratio of total catchment discharge to total storm rainfall). In connection to this, Shuttleworth, Evans, Pilkington, et al. (2019) found no statistically significant change in runoff coefficients following either revegetation or gully blocking, suggesting that the change in the static storage capacity post-treatments was not the primary driver of observed peak reduction and lag increase in these sites.

6.1.2. Kinematic Storage

We define surface "kinematic storage" as the amount of surface water in motion, that is not yet delivered to the outlet, at any given point in time. Contrary to static storage, water in the kinematic storage unit will all eventually leave the catchment as discharge, thus will not alter runoff coefficients. In TOPMODEL, kinematic storage is not explicitly represented by a parameter, but it is inversely proportional to the speed of surface flow response, that is, flow celerity parameter (c). Speed of response in this context is the time it takes for perturbations in the kinematic storage unit, which travel overland, to arrive at the catchment outlet. Thus, during rainfall, rougher surfaces (for example due to revegetation or stream blocking) reduce celerities, retain water within the catchment for longer, and lead to accumulation of water in the kinematic storage unit. Here, for a given rainfall event, the rougher the surface, the larger the kinematic storage volume that the catchment can hold.

6.2. Surface Storage Post-Interventions: Role of Storm Size

6.2.1. Static Storage

We first investigate the overall static storage effect, through the coupled parameters $S_{rz_{max}}+E_p$. Investigation of Figure 6g suggests that in both RV and RG, changes to static storage capacity (i.e., $S_{rz_{max}}+E_p$) are the primary cause of peak reduction in smaller storms ($\leq$0.01 m^3/s), although closely followed by celerity c. In larger storms (>0.01 m^3/s) static storage is much less important for reducing the peaks, but is never negligible (it is at least 10% of the effect of celerity on average). Figure 6i shows that in both sites and in smaller storms, the role of static storage in delaying the flow is small but not negligible. Also static storage capacity in RG delays the flow more than RV in smaller storms. In larger storms however, regardless of treatment, the impact of static storage capacity on peak timing delay is negligible.

Regarding the relative roles of $S_{rz_{max}}$ and E_p following each treatment, their coupling in TOPMODEL (see Section 5.3) means that: (1) if E_p is small relative to the input rainfall intensity, regardless of $S_{rz_{max}}$, evapotranspirative losses are small "during" a storm. If $S_{rz_{max}}$ is also small, there are limited additional evapotranspirative losses between storms because there is little stored water to lose. But if $S_{rz_{max}}$ is large, these losses will extend to the inter-storm period, making more rootzone storage available at the onset of the next storm. In this case the significance of this additional static storage capacity depends on the storm sequencing pattern. (2) if E_p rate is comparable to the input rainfall intensity, regardless of $S_{rz_{max}}$, evapotranspirative losses "during" the storm will also be considerable. If $S_{rz_{max}}$ is small, again these losses primarily occur during the rainfall event but this time they are considerable. If $S_{rz_{max}}$ is also large however, the evapotranspirative losses will extend to the inter-storm period and at higher rates, leading to even more rootzone storage available at the onset of the succeeding storm which must be filled before surface runoff can be generated (i.e., more peak reduction and delay).

In this context, Figure 6c shows that the increase in $S_{rz_{max}}$ due to revegetation has only a small impact on peak discharge independent of storm size. This suggests that revegetation changes static storage capacity primarily by raising E_p rates emptying the store more rapidly rather than by increasing interception storage capacity. This in turn suggests that the mechanism through which revegetation increases the catchment's static storage capacity is the evapotranspiration occurring "during" the storm. Further comparing RV and RG in panel (g) shows that independent of storm size, on average the net effect of $S_{rz_{max}}+E_p$ in peak reduction is not significantly different. On the other hand, Figure 6e shows that gully blocking increases not only E_p (as it did for RV) but also $S_{rz_{max}}$. Together these two observations suggest that, gully blocking has increased $S_{rz_{max}}$, but has reduced E_p rates such that blocks have no net effect on static storage capacity when compared to the revegetation-only case. This in turn suggests that the E_p rates of a unit vegetated surface area are higher than that of a pond behind the blocks.

Higher E_p rates of a unit vegetated surface area than that of a pond predicted by our model is surprising, but is consistent with findings from a range of other studies which examine the ratio of actual evapotranspiration E_a to open water evaporation E_w for wetland ecosystems in a range of climates (e.g., Campbell &

Williamson, 1997; Dolan et al., 1984; Drexler et al., 2004). These studies suggest significant variability in the ratio, both above and below 1 that is strongly sensitive to biophysical factors linked to vegetation type and water supply. In particular, when canopy resistance to evaporation is low (<100 s/m) wetland vegetation increases E_a relative to open water (Mohamed et al., 2012); similarly, evaporation from wet leaves may exceed that of open water because the vegetation is able to withdraw more energy from the atmospheric boundary layer by reducing its aerodynamic resistance to evaporation (Savenije, 2004).

In terms of peak timing, Figure 6i suggests that in large storms and independent of treatment, the delay caused by the process of filling static storage is negligible. In smaller storm this delay is more significant and is on average 5 min in RV and 10 min in RG. These suggest that the additional static storage capacity added by revegetation can slightly delay the flow but only in smaller storms; it also suggest that gully blocking further increases this delay in smaller storms, but it is still not significant in large storms. Additional inference can be made by comparing panels (d and f) in Figure 6, which suggest that the underlying mechanisms through which these treatments delay the flow are different; in RV these lag times are primarily delivered through increased E_p rates whereas in RG primarily through increased $S_{rz_{max}}$. Again, this suggests increased static storage capacity as a result of gully blocking but reduced E_p relative to revegetation.

6.2.2. Kinematic Storage

For both RV and RG, averaged across all storm sizes, celerity is the most important parameter contributing to both peak discharge reduction and lag time increase (Figures 6h and 6j). Similar to the findings of Shuttleworth, Evans, Pilkington, et al. (2019), the fraction of the hydrograph change that can be accomplished by changing celerity alone at RG is almost double that at RV, both in peak discharge reduction and lag increase. Note that being semi-distributed, TOPMODEL only treats the hillslopes, channels and blocks in a bulk sense and therefore the celerity in the model is also a bulk celerity averaged over the catchment area. Thus, although it will not represent the true celerity of different parts of the hillslope, channel or individual blocks, it does represent the average response-speed of surface flow necessary to reproduce the observed hydrograph for the studied rainfall-catchment combinations. Therefore, the significant reduction in celerity predicted by the model following these interventions has three important implications. First, the additional kinematic storage volume due to both interventions (RV and RG) is consistently the dominant process for both peak reduction and lag increase. This is important because it suggests that hydrograph attenuation (magnitude reduction and delay due to changes in hydrograph shape rather than volume) resulting from interventions is not limited to smaller storms. Second, the kinematic storage volume associated with gully blocks is much more important than their static storage capacity. Third, in terms of kinematic storage, the additional impact of gully blocking is as large as that of revegetation (see Figures 6g–6j).

Although it is intuitive that gully blocking should significantly reduce mean particle velocities of surface flow (as measured by tracers), it is far less clear how gully blocking reduces celerity (see Section 5.4). For example, celerity estimates based on the St. Venant equation relate c to both flow depth and velocity ($c = v \pm \sqrt{(gy)}$, $v[LT^{-1}]$: velocity, $g[LT^{-2}]$: gravitational constant, $y[L]$ flow depth) (McDonnell & Beven, 2014). It is easier to imagine how celerity might be reduced in the case of permeable blocks due to their nonlinear storage discharge relationship (Metcalfe, Beven, Hankin, & Lamb, 2017; Milledge et al., 2015). For an impermeable block such as a timber dam, the situation is more complex. Flow depth changes discontinuously along the channel length: due to gully bed slope, flow depth gradually increases toward its maximum at the block, but then rapidly decreases to the thickness of the sheet of water flowing over the top of the block. Therefore, compared to a non-blocked case, mean particle velocity is much lower behind the block, while the $\sqrt{(gy)}$ term in celerity is much higher due to deeper flow. However, in the thin flow over the block the $\sqrt{(gy)}$ term suddenly and significantly decreases, while at the same time mass continuity forces a rapid increase in mean particle velocity (increasing the v term). In reality water then flows into the subsequent blocks, with block interactions (block spacing, configuration, etc.) introducing additional complexity.

In our study site (RG) gullies were blocked with 20 impermeable timber dams and 17 cobble dams that were initially permeable but that tend to clog with sediments reducing their permeability over time. Our model results suggest considerable celerity reduction due to gully blocking, but the mechanism for such a reduction as well as the extent to which it can be attributed to permeable or impermeable blocks cannot be
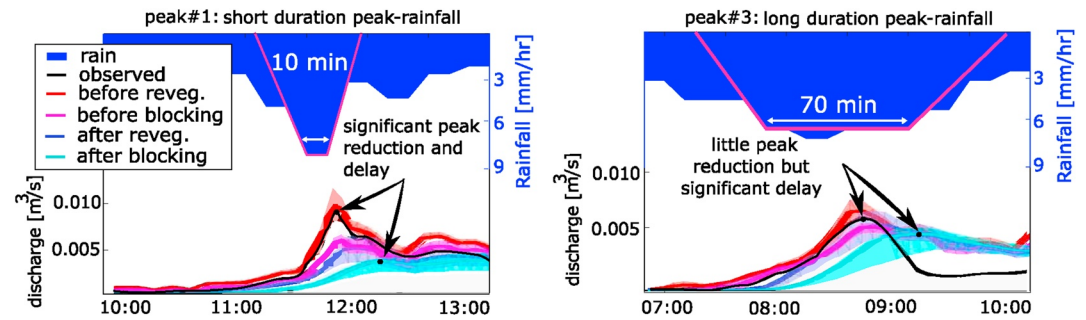
Figure 7. Comparing duration of peak rainfall intensity and its impact on peak magnitude reduction.

untangled in this modeling framework. Explicit surface flow routing and block permeability representation could feasibly address this uncertainty and would be valuable for informing future block design.

6.3. Role of Storm Duration

In the post-intervention period, despite significantly delaying the flow, Peak 3 in Figure 5b exhibits very little reduction in peak discharge when compared with other peaks, and particularly with Peak 1, which is very similar in magnitude and overall hyetograph shape. This can also be seen in the peak discharge PDFs of Peak 3 where there is no separation between "pre" and "post" PDFs for either treatment, meaning that the magnitude of change in discharge was not significantly different (with 95% confidence) between these periods.

The main observed difference between rainfall-runoff behaviors of Peaks 1 and 3 is in their duration of peak rainfall intensity (see Figure 7). The storm responsible for Peak 3 has a hyetograph that features a wide peak, such that the maximum intensity is almost constant for a span of 70 min. The hyetograph associated with Peak 1 has a much narrower peak, spanning only around 10 min. On the other hand, the estimated post-restoration mean lag times for RV and RG were 28 and 50 min, respectively. Thus, the duration of the peak rainfall intensity in Peak 1 is shorter than the mean lag times provided by both sites, whereas for Peak 3 duration of peak intensity rainfall exceeds their lag times.

During Peak 3 the duration of peak rainfall intensity exceeded the time to equilibrium for the catchment (the time for the floodwave to propagate from the furthest point in the catchment). As a result, the discharge peak is delayed but not reduced in magnitude (with the catchment in an almost "steady-state"). This phenomenon is independent of the magnitude of the rainfall and is driven by duration, which makes it possible for a prolonged storm event to fill the entire catchment's storage capacity (both static storage and kinematic storage), whilst experiencing no reduction in its magnitude, and thus removing the peak discharge reduction effect of NFM interventions (although not their delaying effect).

Note that Peak 3 has a peak discharge less than half that of the largest storm in the study period, and is only used here to draw attention to the possible impact of such long-duration storms, irrespective of their probability of occurrence. In reality, it may be that the highest intensity storms rarely, or never assume such broad hyetograph shapes. Constraining this would require an Intensity-Duration-Frequency (IDF) study. Nonetheless, our findings and those of Gao, Kirkby, and Holden (2018), suggest that the effectiveness of NFM interventions can depend on hyetograph characteristics.

6.4. NFM Implications at Catchment-Scale

Revegetation and gully blocking activities in our study catchments were motivated by, and designed for moorland restoration rather than for NFM. It is unlikely that a flood management focus would have substantially changed the revegetation treatment; however, it likely would have changed the gully block designs/configurations, which were introduced to elevate water tables, trap sediment and prevent further gully erosion, rather than to attenuate flow. Given the difference in focus, it is perhaps surprising how successful the interventions have been in reducing and delaying peak discharge.

Further, our results show that the percentage reduction in peak discharge from a micro-catchment can vary widely (5%–90% in Figure 6a), and that this may be due to rainfall characteristics (Section 6.3). If so, the interventions studied here will be least effective in reducing micro-catchment peak discharge for long-duration, "frontal" rainfall, and more effective for shorter "convectional" rainfall, such as that responsible for the floods in nearby Glossop (August 2002).

It is also important to note that regardless of the duration of peak rainfall intensity, lag times are always considerably increased following these interventions (minimum 10 min in RV and 30 min in RG). Consequently, even in the longest frontal rainstorms, these increased lag times for upland headwater subcatchments can be utilized to attenuate the downstream flood wave, as the delayed delivery of water will allow flood peaks from other downstream subcatchments to pass first, before delivering water to catchments draining to communities at risk (Pattison et al., 2014). This, however, comes with the caveat that when alterations are primarily to flood wave timing rather than magnitude, it is possible that interventions synchronize subcatchments rather than bringing them out of synch. To avoid such situations, catchment-scale models (such as Metcalfe, Beven, Hankin, & Lamb, 2018) can be useful in testing different spatial patterns of NFM interventions and their impact on flood risk throughout a river network.

Nonetheless, possible susceptibility of these interventions to long-duration, frontal rainfall suggests that: revegetation and gully blocking of the design examined here, may not provide full protection against flooding in some cases. If so, this would be because the catchment's total surface storage (static + kinematic storage) was not sufficiently increased by these interventions to reduce peak magnitude even in longest duration events. Note that the restoration interventions' main objective is to raise the water tables, which translates into reduced available subsurface storage. Therefore, to extend these interventions' effectiveness to long frontal rainstorms while remaining within the restoration objectives, the catchment's total "surface" storage needs to be increased. We propose two approaches.

First, trade-off static storage capacity for more kinematic storage volume. Because in our largest storms, total static storage capacity was insufficient to significantly reduce peak magnitude or increase lag times. Instead, celerity reduction had a far greater effect. Thus making the blocks "optimally" permeable could further increase surface roughness and reduce celerity. This would better utilize the kinematic storage effect, which might not currently be maximally exploited. This can be done, for instance, by inserting pipes, or cutting holes into otherwise impermeable gully blocks (e.g., Metcalfe, Beven, Hankin, & Lamb, 2017; Milledge et al., 2015), such that surface flow is partitioned into storage during times of peak flow. Here, for a given storm, optimal design would result in blocks that: fill to capacity, do so only at peak flow, and never overtop.

Second, increase the total static storage capacity (as also suggested by Holden, Moody, et al. [2018]) by creating more shallow open water pools in other parts of the catchment as well as in-channels and gullies, such that its volume is comparable to the large water volumes associated with flood relevant storms. Note that this may also increase kinematic storage volume by further reducing surface flow celerities, but comes with the caveat that since total storm-water volumes scale with catchment area, full catchment-scale implementation may be costly/unfeasible.

7. Limitations & Future Research

In this study we have sought to rigorously account for parametric and data uncertainties. However, model structural uncertainties remain and are difficult to quantify. These structural uncertainties emerge because there are other interesting and potentially important processes that may be influencing rainfall-runoff at time-scales relevant to our study, but which we do not represent explicitly. Of particular relevance are, peat swelling in response to rainfall leading to an "elastic storativity" (Nijp et al., 2019) effect (a member of the family of echo-hydrological self-regulating properties of peat; see Waddington et al., 2015), the known depth-dependence of both velocity (Holden, Kirkby, et al., 2008) and celerity (McDonnell & Beven, 2014), and possibility of (semi)permeable gully blocks with a storage discharge relationship.

Peat swelling can result in changes to (a) subsurface storage capacity, and (b) surface topography, which in turn partitions into (b1) changes to depression storage, equivalent to changes to our static storage, and (b2) changes to runoff pathways (rather than skin friction), equivalent to changes to our kinematic storage.

However, while we do not explicitly account for peat surface elevation change due to peat expansion/contraction, our model implicitly accounts for such process by estimating the water table depth relative the peat surface. This has the advantage that it avoids introducing another uncertain parameter in the model. As a result, although the model does not individually represent peat surface elevation or water table height, it captures the "net" effect of these two together in a lumped way. Regarding (b1), changes to depression storage would be reflected in the static storage capacity, which represents a bulk interception storage in the model; hence (b1) is also accounted for in lumped manner. Regarding (b2), our study sites have average flow path gradients of roughly 10°, whereas peat swelling can only result in subtle (<1.4°, with <0.1° more dominant) changes to surface topography (Nijp et al., 2019). Consequently, in this case, impact of peat-deformation on kinematic storage volume is likely to be minor relative to those due to skin friction; such that our model's kinematic storage predictions are still representative of the impact of vegetation/gully blocking as surface roughness elements.

Treating celerity as independent of flow depth/thickness, could result in underestimation of the kinematic storage in some parts of the catchment, and over-estimation in others; depending on whether the depth-averaged celerity was more or less than the true depth-dependent values in those parts. When calculating kinematic storage volumes, it is conceivable that some of the errors associated with depth-averaging the celerity would cancel out due to smoothing that occurs by averaging over- and under-estimated values across different parts of the catchment; but insignificance of neglecting such process is not a given.

Also, our current model structure does not explicitly account for potential differences in gully block permeability between wooden and cobble blocks. Permeable blocks would convert some static storage to kinematic storage during the storm, the amount of which is controlled by permeability levels. Based on field visits in our sites, most cobble stone dams have likely been clogged up with sediments. However, the precise state of (im)permeability of the cobble dams during our study periods is unknown. As such, although it seems less likely that including such process in our model would have considerably altered our static and/or kinematic storage estimates, but it cannot be said with confidence.

Based on the above descriptions, it is possible, though not obvious, that explicitly including all of the above processes could provide useful insights. What is obvious, however, is that our static and kinematic storage predictions lump different catchment processes within them. To ungroup and untangle these processes from our lumped predictions, additional model complexity is required. But representing each additional process would require a step up in model complexity and the resulting expansion of its parameterization. We note that hydrological models with more than 5 uncertain parameters are rarely justified (see for example Perrin et al. [2001]), due to the limited information content of the calibration data. This is not to say that additional processes are not worth including, but that there is a trade-off between model complexity and "identifiability" of its parameters through calibration, which (1) calls for future attempts to also focus on integrating the available, empirically based catchment functions into model structures, but with minimal parametric overhead, to help untangle and ungroup the lumped model predictions; and (2) highlights the importance of prioritizing the inclusion of hydrological processes in a catchment-specific way.

Because the degree to which interpretations of lumped model predictions provide useful information, depends on how well the catchment is expected to conform to the lumped representation being used. Our study is concerned with steeply sloping, blanket peat catchments (characteristic of the dominant UK peat type), which are dominated by topography-driven, saturation-excess overland flow (Evans, Burt, et al., 1999). For this reason, we expect conceptualization and modeling of our catchments based on subsurface storage, and static and kinetic surface storage to adequately capture the most dominant processes in this case. However, following our approach, future research could usefully attempt to estimate the impact magnitude of each individual process that may be hidden within our lumped estimations.

Finally, as was discussed in Section 2, our 2012 storm (panel c in Figure 5) had an intensity-duration combination that has previously caused flooding in the region. But, we also know that none of the storms in our records resulted in flooding downstream of our micro-catchments (they resulted in downstream flood peaks up to 79% of that associated with flooding in the nearby region). This highlights the complex dependency of downstream discharge on catchment properties, antecedent conditions, and rainfall characteristics.

We note that based on our results in Figure 6, above the discharge threshold of 1 m³/s, there was clear division between the role of static and kinematic storage in reducing and delaying the floodpeaks. Thus, we are confident that presence of larger storms (that resulted in flooding) would not have changed the underlying processes that we identify as responsible for delivering NFM benefits. But the absolute impact magnitudes of these processes are likely dependent on the storm events, distributed spatial configuration of the interventions, and catchment properties/area. Thus, our impact magnitude estimations are not necessarily applicable to the same interventions but for different spatial scales, configurations or events. Constraining this would require applying our model to a large spatial scale (at which damaging flood can occur), with a large number of storms and with a range of sequencing patterns. We defer this to a separate study.

8. Conclusions

We have shown that a rainfall-runoff model can be inverted to corroborate findings and expand process understanding associated with BACI observations. Confidence in our model predictions comes from: (1) a large number of behavioral parameter-sets (>600) at all site-intervention combinations; (2) good agreement between predicted and observed discharge for behavioral models both in terms of overall fit (NSE) and peak characteristics (magnitude and timing); (3) agreement between predicted and observed discharge that is retained with little degradation in performance metrics for a test period held back during calibration; and (4) agreement with findings from the BACI experiment in terms of the distribution of changes in magnitude and timing of peak discharge.

In summary, we find that revegetation and gully blocking activities motivated by, and designed for moorland restoration also provide significant additional catchment surface storage (static + kinematic storage), which reduces peak discharge and increases lag times, independent of storm size. Revegetation increases static storage capacity through more evapotranspiration almost exclusively "during" storms, and not through increased rainfall interception. Gully blocking extends the evapotranspirative losses to the inter-storm period through increasing the static storage capacity, but also reduces catchment average evapotranspiration rates. These two effects approximately offset one another, such that the overall impact of blocking on static storage capacity is negligible when compared to revegetation alone. Importantly though, for flood relevant storms, static storage plays only a secondary role to kinematic storage. In these storms, independent of intervention type, the key catchment changes responsible for reducing peak discharge and delaying lag times is increased surface roughness through its impact on surface flow celerity.

The dominant role of roughness on discharge attenuation means that hyetograph shape can strongly influence the interventions′ effectiveness in reducing peak discharge. Long, broad, rainfall intensity peaks can lead to near steady-state discharge responses from these small catchments; in which case, reduced surface flow celerities still increase lag times but they no longer reduce peak discharge. Although both are important, the latter has a more reliable impact on downstream discharge because lag time increase in headwater catchments generally attenuate downstream discharge but can amplify it by synchronizing subcatchment hydrographs. This highlights the importance of catchment-scale modeling to test the impact of intervention scenarios before implementation. Our pre- and post-intervention parameter-sets may provide useful constraints on efforts to parameterize such models.

Interventions focused more sharply on NFM could both reduce peak discharge and delay peak timing by increasing total catchment surface static and kinematic storage. We propose two approaches: (1) trade-off static storage capacity for more kinematic storage volume by making the blocks "optimally" permeable; (2) increase the total static storage capacity by creating more shallow open water pools in other parts of the catchment as well as in-channels and gullies.

Finally, while this field-scale numerical study contributes to the evidence-base for NFM′s effectiveness, it also provides a basis for modeling these interventions in the future. But catchment-scale model applications are required to: extend our findings to spatial scales at which flooding can cause socioeconomic damage, and to provide a tool for optimizing the distributed configuration of these interventions.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The hydrological data were collected as part of the Making Space for Water projects funded by DEFRA and the UK Environment Agency. The work benefited from feedback from the wider PROTECT project team. The hydrological data used in this research are included in this study: Shuttleworth, Evans, Pilkington, et al. (2019).

References

Acreman, M., & Holden, J. (2013). How wetlands affect floods. *Wetlands*, *33*(5), 773–786. https://doi.org/10.1007/s13157-013-0473-2

Armstrong, A., Holden, J., Kay, P., Francis, B., Foulger, M., Gledhill, S., & Walker, A. (2010). The impact of peatland drain-blocking on dissolved organic carbon loss and discolouration of water; results from a national survey. *Journal of Hydrology*, *381*(1–2), 112–120. https://doi.org/10.1016/j.jhydrol.2009.11.031

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K., & Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K. J. (2020). A history of the concept of time of concentration. *Hydrology and Earth System Sciences*, *24*(5), 2655–2670. https://doi.org/10.5194/hess-24-2655-2020

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Bulletin*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2007wr006726

Buendia, C., Bussi, G., Tuset, J., Vericat, D., Sabater, S., Palau, A., & Batalla, R. J. (2016). Effects of afforestation on runoff and sediment load in an upland Mediterranean catchment. *Science of the Total Environment*, *540*, 144–157. https://doi.org/10.1016/j.scitotenv.2015.07.005

Bussi, G., Rodríguez-Lloveras, X., Francés, F., Benito, G., Sánchez-Moya, Y., & Sopeña, A. (2013). Sediment yield model implementation based on check dam infill stratigraphy in a semiarid Mediterranean catchment. *Hydrology and Earth System Sciences*, *17*(8), 3339–3354. https://doi.org/10.5194/hess-17-3339-2013

Calder, I. R., Harding, R. J., & Rosier, P. T. W. (1983). An objective assessment of soil-moisture deficit models. *Journal of Hydrology*, *60*(1–4), 329–355. https://doi.org/10.1016/0022-1694(83)90030-6

Campbell, D. I., & Williamson, J. L. (1997). Evaporation from a raised peat bog. *Journal of Hydrology*, *193*(1–4), 142–160. https://doi.org/10.1016/s0022-1694(96)03149-6

Ciach, G. J. (2003). Local random errors in tipping-bucket rain gauge measurements. *Journal of Atmospheric and Oceanic Technology*, *20*(5), 752–759. https://doi.org/10.1175/1520-0426(2003)20<752:lreitb>2.0.co;2

Cole, B., McMorrow, J., & Evans, M. (2014). Empirical modelling of vegetation abundance from airborne hyperspectral data for upland peatland restoration monitoring. *Remote Sensing*, *6*(1), 716–739. https://doi.org/10.3390/rs6010716

Dadson, S. J., Hall, J. W., Murgatroyd, A., Acreman, M., Bates, P., Beven, K., et al. (2017). A restatement of the natural science evidence concerning catchment-based 'natural' flood management in the UK. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *473*(2199), 20160706. https://doi.org/10.1098/rspa.2016.0706

Daniels, S. M., Agnew, C. T., Allott, T. E. H., & Evans, M. G. (2008). Water table variability and runoff generation in an eroded peatland, South Pennines, UK. *Journal of Hydrology*, *361*(1–2), 214–226. https://doi.org/10.1016/j.jhydrol.2008.07.042

Dixon, S. D., Qassim, S. M., Rowson, J. G., Worrall, F., Evans, M. G., Boothroyd, I. M., & Bonn, A. (2014). Restoration effects on water table depths and CO_2 fluxes from climatically marginal blanket bog. *Biogeochemistry*, *118*(1–3), 159–176. https://doi.org/10.1007/s10533-013-9915-4

Dolan, T. J., Hermann, A. J., Bayley, S. E., & Zoltek, J., Jr. (1984). Evapotranspiration of a Florida, USA, freshwater wetland. *Journal of Hydrology*, *74*(3–4), 355–371. https://doi.org/10.1016/0022-1694(84)90024-6

Drexler, J. Z., Snyder, R. L., Spano, D., & Paw U, K. T. (2004). A review of models and micrometeorological methods used to estimate wetland evapotranspiration. *Hydrological Processes*, *18*(11), 2071–2101. https://doi.org/10.1002/hyp.1462

Evans, M., Allott, T., Holden, J., Flitcroft, C., & Bonn, A. (2005). *Understanding gully blocking in deep peat*. Derbyshire: Moors for the Future Report (Vol. 4).

Evans, M. G., Burt, T. P., Holden, J., & Adamson, J. K. (1999). Runoff generation and water table fluctuations in blanket peat: Evidence from UK data spanning the dry summer of 1995. *Journal of Hydrology*, *221*(3–4), 141–160. https://doi.org/10.1016/s0022-1694(99)00085-2

Evans, M., & Warburton, J. (2011). *Geomorphology of upland peat: Erosion, form and landscape change*. John Wiley & Sons.

Evans, M., Warburton, J., & Yang, J. (2006). Eroding blanket peat catchments: Global and local implications of upland organic sediment budgets. *Geomorphology*, *79*(1–2), 45–57. https://doi.org/10.1016/j.geomorph.2005.09.015

Féral, L., Sauvageot, H., Castanet, L., Lemorton, J., Cornet, F., & Leconte, K. (2006). Large-scale modeling of rain fields from a rain cell deterministic model. *Radio Science*, *41*(02), 1–21. https://doi.org/10.1029/2005rs003312

Freer, J., Beven, K. J., & Peters, N. (2003). Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure. *Calibration of Watershed Models*, *69*, 88.

Gao, J., Holden, J., & Kirkby, M. (2015). A distributed TOPMODEL for modelling impacts of land-cover change on river flow in upland peatland catchments. *Hydrological Processes*, *29*(13), 2867–2879. https://doi.org/10.1002/hyp.10408

Gao, J., Holden, J., & Kirkby, M. (2016). The impact of land-cover change on flood peaks in peatland basins. *Water Resources Research*, *52*(5), 3477–3492. https://doi.org/10.1002/2015wr017667

Gao, J., Holden, J., & Kirkby, M. (2017). Modelling impacts of agricultural practice on flood peaks in upland catchments: An application of the distributed TOPMODEL. *Hydrological Processes*, *31*(23), 4206–4216. https://doi.org/10.1002/hyp.11355

Gao, J., Kirkby, M., & Holden, J. (2018). The effect of interactions between rainfall patterns and land-cover change on flood peaks in upland peatlands. *Journal of Hydrology*, *567*, 546–559. https://doi.org/10.1016/j.jhydrol.2018.10.039

Grayson, R., Holden, J., & Rose, R. (2010). Long-term change in storm hydrographs in response to peatland vegetation change. *Journal of Hydrology*, *389*(3–4), 336–343. https://doi.org/10.1016/j.jhydrol.2010.06.012

Holden, J. (2005). Peatland hydrology and carbon release: Why small-scale process matters. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *363*(1837), 2891–2913. https://doi.org/10.1098/rsta.2005.1671

Holden, J. (2009). Peatland hydrology and carbon release. In A. Bonn, T. Allott, K. Hubacek, J. Stewart (Eds.), *Drivers of environmental change in uplands* (pp. 113–134). Routledge.

Holden, J., & Burt, T. P. (2002). Infiltration, runoff and sediment production in blanket peat catchments: Implications of field rainfall simulation experiments. *Hydrological Processes*, *16*(13), 2537–2557. https://doi.org/10.1002/hyp.1014

Holden, J., & Burt, T. P. (2003). Runoff production in blanket peat covered catchments. *Water Resources Research*, *39*(7), 1191–1200. https://doi.org/10.1029/2002wr001956

Holden, J., Chapman, P. J., & Labadz, J. C. (2004). Artificial drainage of peatlands: Hydrological and hydrochemical process and wetland restoration. *Progress in Physical Geography*, *28*(1), 95–123. https://doi.org/10.1191/0309133304pp403ra

Holden, J., Kirkby, M. J., Lane, S. N., Milledge, D. G., Brookes, C. J., Holden, V., & McDonald, A. T. (2008). Overland flow velocity and roughness properties in peatlands. *Water Resources Research*, *44*(6). https://doi.org/10.1029/2007wr006052

Holden, J., Moody, C. S., Edward Turner, T., McKenzie, R., Baird, A. J., Billett, M. F., et al. (2018). Water-level dynamics in natural and artificial pools in blanket peatlands. *Hydrological Processes*, *32*(4), 550–561. https://doi.org/10.1002/hyp.11438

Jones, M., Fowler, H. J., Kilsby, C. G., & Blenkinsop, S. (2010). An updated regional frequency analysis of United Kingdom extreme rainfall 1961–2009. *Proceedings of BHS Conference* (pp. 110–118).

Kavetski, D., Kuczera, G., & Franks, S. W. (2003). Semidistributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC. *Water Resources Research*, *39*(9). https://doi.org/10.1029/2003wr002122

Keeland, B. D., Dowd, J. F., & Hardegree, W. S. (1997). Use of inexpensive pressure transducers for measuring water levels in wells. *Wetlands Ecology and Management*, *5*(2), 121–129. https://doi.org/10.1023/a:1008203625917

Kirkby, M. J. (1997). TOPMODEL: A personal view. *Hydrological Processes*, *11*(9), 1087–1097. https://doi.org/10.1002/(sici)1099-1085(199707)11:9<1087::aid-hyp546>3.0.co;2-p

Lane, S. N., Brookes, C. J., Kirkby, M. J., & Holden, J. (2004). A network-index-based version of TOPMODEL for use with high-resolution digital topographic data. *Hydrological Processes*, *18*(1), 191–201. https://doi.org/10.1002/hyp.5208

Lane, S. N., & Milledge, D. G. (2013). Impacts of upland open drains upon runoff generation: A numerical assessment of catchment-scale impacts. *Hydrological Processes*, *27*(12), 1701–1726. https://doi.org/10.1002/hyp.9285

Lindsay, R., Charman, D. J., Everingham, F., O'reilly, R. M., Palmer, M. A., Rowell, T. A., & Stroud, D. A. (1988). *The flow country: The peatlands of Caithness and Sutherland*. Joint Nature Conservation Committee.

McDonnell, J. J., & Beven, K. (2014). Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph. *Water Resources Research*, *50*(6), 5342–5350. https://doi.org/10.1002/2013wr015141

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

Medici, C., Butturini, A., Bernal, S., Vázquez, E., Sabater, F., Vélez, J. I., & Francés, F. (2008). Modelling the non-linear hydrological behaviour of a small Mediterranean forested catchment. *Hydrological Processes: An International Journal*, *22*(18), 3814–3828. https://doi.org/10.1002/hyp.6991

Metcalfe, P., Beven, K., & Freer, J. (2015). Dynamic TOPMODEL: A new implementation in R and its sensitivity to time and space steps. *Environmental Modelling & Software*, *72*, 155–172. https://doi.org/10.1016/j.envsoft.2015.06.010

Metcalfe, P., Beven, K., Hankin, B., & Lamb, R. (2017). A modelling framework for evaluation of the hydrological impacts of nature-based approaches to flood risk management, with application to in-channel interventions across a 29-km^2 scale catchment in the United Kingdom. *Hydrological Processes*, *31*(9), 1734–1748. https://doi.org/10.1002/hyp.11140

Metcalfe, P., Beven, K., Hankin, B., & Lamb, R. (2018). A new method, with application, for analysis of the impacts on flood risk of widely distributed enhanced hillslope storage. *Hydrology and Earth System Sciences*, *22*(4), 2589–2605. https://doi.org/10.5194/hess-22-2589-2018

Milledge, D., Odoni, N., Allott, T., Evans, M., Pilkington, M., Walker, J., et al. (2015). Annex 6. Flood risk modelling. In M. Pilkington, J. Walker, R. Maskill, T. Allott, M. Evans (Eds.), *Restoration of Blanket bogs; flood risk reduction and other ecosystem benefits*. Edale: Moors for the Future Partnership.

Mohamed, Y. A., Bastiaanssen, W. G. M., Savenije, H. H. G., Van den Hurk, B. J. J. M., & Finlayson, C. M. (2012). Wetland versus open water evaporation: An analysis and literature review. *Physics and Chemistry of the Earth, Parts A/B/C*, *47–48*, 114–121. https://doi.org/10.1016/j.pce.2011.08.005

Nijp, J. J., Metselaar, K., Limpens, J., Bartholomeus, H. M., Nilsson, M. B., Berendse, F., & van der Zee, S. E. (2019). High-resolution peat volume change in a northern peatland: Spatial variability, main drivers, and impact on ecohydrology. *Ecohydrology*, *12*(6), e2114. https://doi.org/10.1002/eco.2114

Parry, L. E., Holden, J., & Chapman, P. J. (2014). Restoration of blanket peatlands. *Journal of Environmental Management*, *133*, 193–205. https://doi.org/10.1016/j.jenvman.2013.11.033

Pattison, I., Lane, S. N., Hardy, R. J., & Reaney, S. M. (2014). The role of tributary relative timing and sequencing in controlling large floods. *Water Resources Research*, *50*(7), 5444–5458. https://doi.org/10.1002/2013wr014067

Peña, L. E., Barrios, M., & Francés, F. (2016). Flood quantiles scaling with upper soil hydraulic properties for different land uses at catchment scale. *Journal of Hydrology*, *541*, 1258–1272. https://doi.org/10.1016/j.jhydrol.2016.08.031

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, *242*(3–4), 275–301. https://doi.org/10.1016/s0022-1694(00)00393-0

Petersen-Øverleir, A., & Reitan, T. (2005). Uncertainty in flood discharges from urban and small rural catchments due to inaccurate head measurement. *Hydrology Research*, *36*(3), 245–257. https://doi.org/10.2166/nh.2005.0018

Pollock, M. D., O'Donnell, G., Quinn, P., Dutton, M., Black, A., Wilkinson, M. E., et al. (2018). Quantifying and mitigating wind-induced undercatch in rainfall measurements. *Water Resources Research*, *54*(6), 3863–3875. https://doi.org/10.1029/2017wr022421

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2009wr008328

Rodda, J. C. (1967). The rainfall measurement problem. Proceedings of IAHS General Assembly (Vol. 78, pp. 215–231). Bern. IAHS Publication.

Rothwell, J. J., Evans, M. G., & Allott, T. E. H. (2007). Lead contamination of fluvial sediments in an eroding blanket peat catchment. *Applied Geochemistry*, *22*(2), 446–459. https://doi.org/10.1016/j.apgeochem.2006.11.002

Ruiz-Pérez, G., Medici, C., Latron, J., Llorens, P., Gallart, F., & Francés, F. (2016). Investigating the behaviour of a small Mediterranean catchment using three different hydrological models as hypotheses. *Hydrological Processes*, *30*(13), 2050–2062. https://doi.org/10.1002/hyp.10738

Sauvageot, H., Mesnard, F., & Tenório, R. S. (1999). The relation between the area-average rain rate and the rain cell size distribution parameters. *Journal of the Atmospheric Sciences*, *56*(1), 57–70. https://doi.org/10.1175/1520-0469(1999)056<0057:trbtaa>2.0.co;2

Savenije, H. H. G. (2004). The importance of interception and why we should delete the term evapotranspiration from our vocabulary. *Hydrological Processes*, *18*(8), 1507–1511. https://doi.org/10.1002/hyp.5563

Shen, J. (1981). *Discharge characteristics of triangular-notch thin-plate weirs (No. 1617)*. United States Department of the Interior, Geological Survey.

Shuttleworth, E. L., Evans, M. G., Hutchinson, S. M., & Rothwell, J. J. (2015). Peatland restoration: controls on sediment production and reductions in carbon and pollutant export. *Earth Surface Processes and Landforms*, *40*(4), 459–472. https://doi.org/10.1002/esp.3645

Shuttleworth, E. L., Evans, M. G., Pilkington, M., Spencer, T., Walker, J., Milledge, D., & Allott, T. E. H. (2019). Restoration of blanket peat moorland delays stormflow from hillslopes and reduces peak discharge. *Journal of Hydrology X*, *2*, 100006. https://doi.org/10.1016/j.hydroa.2018.100006

Stocker, B. D., Spahni, R., & Joos, F. (2014). DYPTOP: A cost-efficient TOPMODEL implementation to simulate sub-grid spatio-temporal dynamics of global wetlands and peatlands. *Geoscientific Model Development*, *7*(6), 3089–3110. https://doi.org/10.5194/gmd-7-3089-2014

Van Straten, G. T., & Keesman, K. J. (1991). Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example. *Journal of Forecasting*, *10*(1–2), 163–190. https://doi.org/10.1002/for.3980100110

Vrugt, J. A., & Beven, K. J. (2018). Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM(LOA) algorithm. *Journal of Hydrology*, *559*, 954–971. https://doi.org/10.1016/j.jhydrol.2018.02.026

Waddington, J. M., Morris, P. J., Kettridge, N., Granath, G., Thompson, D. K., & Moore, P. A. (2015). Hydrological feedbacks in northern peatlands. *Ecohydrology*, *8*(1), 113–127. https://doi.org/10.1002/eco.1493

Wallage, Z. E., Holden, J., & McDonald, A. T. (2006). Drain blocking: An effective treatment for reducing dissolved organic carbon loss and water discolouration in a drained peatland. *Science of the Total Environment*, *367*(2–3), 811–821. https://doi.org/10.1016/j.scitotenv.2006.02.010

Xu, J., Morris, P. J., Liu, J., & Holden, J. (2018). PEATMAP: Refining estimates of global peatland distribution based on a meta-analysis. *Catena*, *160*, 134–140. https://doi.org/10.1016/j.catena.2017.09.010