Numerical daemons of hydrological models are summoned by extreme precipitation

Peter T. La Follette[1], Adriaan J. Teuling[1], Nans Addor[2], Martyn Clark[3], Koen Jansen[1], and Lieke A. Melsen[1]

[1]Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, Netherlands
[2]Geography, College of Life and Environmental Sciences, University of Exeter, Exeter, UK
[3]Coldwater Laboratory, University of Saskatchewan, Canmore, Alberta, Canada

Correspondence: Adriaan J. Teuling (ryan.teuling@wur.nl)

Abstract. Hydrological models are usually systems of nonlinear differential equations for which no analytical solutions exist and thus rely on approximate numerical solutions. While some studies have investigated the relationship between numerical method choice and model error, the extent to which extreme precipitation like that observed during hurricanes Harvey and Katrina impacts numerical error of hydrological models is still unknown. This knowledge is relevant in light of climate change, where many regions will likely experience more intense precipitation events. In this experiment, a large number of hydrographs is generated with the modular modeling framework FUSE, using eight numerical techniques across a variety of forcing datasets. Multiple model structures, parameter sets, and initial conditions are incorporated for generality. The computational expense and numerical error associated with each hydrograph were recorded. It was found that numerical error (root mean square error) usually increases with precipitation intensity and decreases with event duration. Some numerical methods constrain errors much more effectively than others, sometimes by many orders of magnitude. Of the tested numerical methods, a second-order adaptive explicit method is found to be the most efficient because it has both low numerical error and low computational cost. A basic literature review indicates that many popular modeling codes use numerical techniques that were suggested by this experiment to be sub-optimal. We conclude that relatively large numerical errors might be common in current models, and because these will likely become larger as the climate changes, we advocate for the use of low cost, low error numerical methods.

1   Introduction

Computational hydrological models are pieces of code or software that describe the movement and distribution of water within a region. They enjoy frequent use within and outside of academia, addressing a diversity of topics from the determination of catchment characteristics (Kirchner, 2009; Rempe and Dietrich, 2014; Wrede et al., 2015; Melsen et al., 2018) to determining

20   the security of a city's water supply (Paton et al., 2013) to deciding which areas are in danger of flooding (Jasper et al., 2002; Madsen et al., 2014).

Hydrological models usually have state variables, which describe relevant physical quantities in a given region of the model, and fluxes, which describe how the state variables change over time or space. For example, a state variable could be the amount of water in the unsaturated zone of a catchment, and fluxes interacting with that state variable could be evaporation, percolation

25   to the saturated zone, discharge from the catchment, or precipitation, among others. Differential equations are used to describe the relationships between fluxes and state variables. These differential equations are often highly nonlinear, meaning that it is impossible to obtain their exact solutions. However, approximate solutions to these systems of differential equations are possible through a variety of numerical strategies. Therefore, hydrological models contain mathematical relationships between state variables and fluxes that need to be solved numerically (approximately) rather than analytically (exactly).

30   While discharge predictions resulting from hydrological models are often reasonably accurate (Refsgaard and Knudsen, 1996; Addor et al., 2011; Refsgaard, 1997), they are always subject to errors. Total hydrological model error can be decomposed in a few ways: observational, structural, and numerical, among others. Observational errors are differences between real and observed values caused by inaccurate measurements, and structural errors are differences between model results and observed quantities due to the conceptual simplification or misrepresentation of processes in a model compared to reality. Numerical

35   errors are differences between the exact and approximate solutions to the set of equations composing the model that result from the choice of numerical method used to find an approximate solution (Higham, 2002). While many recent efforts in hydrology have advanced observational data quality (Wrede et al., 2015; Kittel et al., 2018) or structural representations of nature (Wrede et al., 2015; Melsen et al., 2018; Prancevic and Kirchner, 2019; Dralle et al., 2018; Coxon et al., 2014; Melsen and Guse, 2019), relatively few studies investigate the effects of numerical choices on model error. Notably, it has been demonstrated that

40   numerical method choice has a large impact on hydrological modeling error (Clark and Kavetski, 2010; Kavetski and Clark, 2010). These papers test the ability of various numerical methods to approximate exact solutions and to predict real discharges, including fixed step and adaptive methods, implicit, explicit, and semi implicit methods, and first and second order methods. Changing these options leads to differences in how fluxes are calculated, and it is shown that certain combinations of these qualities allow for a relatively accurate approximation of the exact solution of a hydrological model.

45   The numerical daemons papers (Clark and Kavetski, 2010; Kavetski and Clark, 2010) provide useful numerical conventions in observational contexts, mainly under general hydroclimatic conditions. As the climate changes and extreme precipitation events become more common (Trenberth, 2011; Meehl et al., 2005; Prein et al., 2017; Huang et al., 2020), one might expect numerical errors to become larger. It is simple to demonstrate that numerical errors could depend on precipitation extremeness, as in Fig. 1. This figure shows discharge hydrographs resulting from a model using two different numerical methods. The
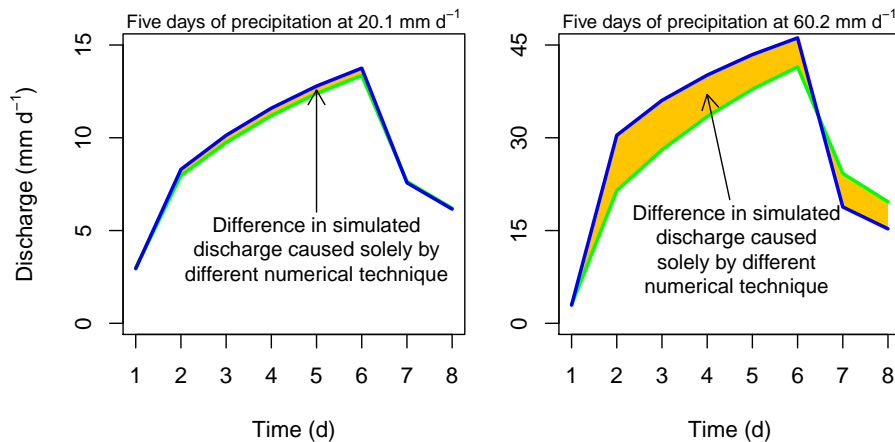
Figure 1. Illustration of the impact of precipitation intensity on differences between discharge simulations from the same model but with two different numerical techniques. Discharge hydrographs are shown resulting from two different numerical methods, indicated by blue and green. These are the fixed step explicit Heun and adaptive implicit Heun methods, respectively, discussed in Section 2. The only difference between the left and right plots is that the precipitation intensity is three times greater during days two through six for the right plot.

50 conditions yielding the left and right plots are identical in every way, except for the precipitation intensity used to simulate discharge, which is three times larger in the graph on the right. The only differences between the blue and green hydrographs are caused by choice of numerical method. Clearly, the two methods agree relatively poorly when higher intensity precipitation data are used; thus there is a greater opportunity for a numerical method to provide an erroneous representation of the exact solution when precipitation is more intense. This is compatible with the findings that models can perform relatively poorly

55 when precipitation becomes intense (Weerts and El Serafy, 2006; Noh et al., 2014; Jasper et al., 2002), where it is possible that numerical error contributes to total modeling error. Further, one might intuitively expect greater numerical errors for more extreme precipitation. The amount of water precipitated by a more extreme storm might be comparable to the total storage of a model, whereas this is not the case for milder storms. Given a coarse enough time step, or a large enough flux, models might simply be overwhelmed. In essence, it is known that numerical error can contribute significantly to total modeling error, and it

60 is known that total modeling errors can increase with increasing precipitation extremeness, but it is currently not known how changing precipitation extremeness (intensity or duration) will impact the numerical error associated with a model.

In this paper, the same eight numerical methods as in Clark and Kavetski (2010) and Kavetski and Clark (2010) are studied (see Section 2 for descriptions of these methods). We aim to determine which qualities of these numerical methods contribute to numerical error as precipitation is varied from mild to extreme. To this end, numerical error associated with each method

65 is assessed over a broad range of precipitation intensities, ranging from mild to more intense than historically observed, for various event durations. To our best knowledge, this is the first instance in which model performance has been assessed at or above world record precipitation (using World Meteorological Organization data) and represents a novel insight into model performance in extreme conditions.

In order to thoroughly investigate the relationship between numerical error and precipitation extremeness for each numerical technique, testing must occur under a variety of conditions. Numerical error could depend on model structure (i.e. the specific relationship between state variables and fluxes, as well as choice of which state variables to include), physical setting (given by a specific combination of model parameters), and antecedent state variable values. This dependence is due to the fact that each of these will affect the set of differential equations composing a model. In order to systematically test these different conditions, a modular modeling framework (MMF) is required. Modular modeling frameworks are tools that are able to rapidly create hydrological models with various processes or structures included, varying numerical schemes (with some MMFs), and other options related to model setup or input. This allows comparison of model results as these settings are systematically altered. The Framework for Understanding Structural Errors (FUSE) is the MMF selected for this experiment.

We are also interested in the change in efficiency of numerical techniques with respect to precipitation extremeness, where efficiency involves both the numerical error associated with a method and its computational expense. Therefore trends in computational expense with changing precipitation are also assessed.

It is possible that many existing hydrological modeling codes use numerical methods that yield large numerical errors. In order to gain insights into how often relatively erroneous numerical methods might be employed, a small literature review is conducted on twelve popular modeling codes, determining the numerical options available in each. When the numerical techniques identified by the literature review are placed in the context of this study, an estimate of the potential magnitude of numerical errors arising from these codes is obtained.

In sum, we use an MMF to study the numerical error and computational expenses associated with different numerical techniques as precipitation varies from mild to extreme. This is carried out under a variety of initial conditions, parameter sets, and model structures for further generality. Trends in error and efficiency are analyzed. Then, a literature review suggests the prevalence of potential numerical errors in popular modeling codes. Finally, we present a discussion of the key findings and their practical implications for modeling, where we hypothesize that numerical errors can be reduced in practice via a careful selection of numerical method.

2   Overview of Numerical Methods used for Approximating Differential Equation Solutions in this Experiment

A hydrological model is usually composed of a system of differential equations, relating state variables and fluxes. These equations are usually highly nonlinear, which means that they cannot be solved in closed form (though this is not always the case; see Coxon et al. (2019)). Thus, a numerical approximation must be used. In this section, a description of the numerical techniques studied in this paper is given. The reader might choose to skip this section if familiar with the terms in bold.

In this experiment, each numerical technique is used to solve a system of coupled ordinary differential equations, where the equations are coupled because the flux from each state variable generally depends on multiple state variables. For example, the net flux from the saturated zone could be a function of the amount of water in the saturated zone and the amount of water in the unsaturated zone. Two options are available for solving these systems: **sequential** solving (or **operator splitting**), where the equations are solved (and the state variables are updated) in some predetermined order, or **simultaneous** solving, where

all equations are solved at once (which requires a **space-state** formulation of the equations which compose a model). FUSE employs simultaneous solving. Sequential solvers are able to use different numerical methods for individual fluxes, which in some fields is desirable (Glowinski et al., 2017). However, when using sequential solvers, the model output may depend on the sequence in which state variables are solved, which is undesirable when comparing different model structures (Clark and Kavetski, 2010; Glowinski et al., 2017). Using the simultaneous method, N equations will be solved at the same time with the same numerical technique, where N is the total number of state variables, so $n = 1, ..., N$ in the equations of this section.

The simplest numerical method studied in this paper is the **fixed step explicit Euler method**. This method works by evaluating the flux for a state variable at the start of a time step, then adding this flux to the state variable at the beginning of the time step to generate the value for the state variable at the end of the time step. An illustrated example is given in Fig. 2a. Symbolically,

$$S_n(t + \Delta t) = S_n(t) + f_n(\boldsymbol{S}(t), t)\Delta t + O(\Delta t) \tag{1}$$

$$S_n(t + \Delta t) \approx S_n(t) + f_n(\boldsymbol{S}(t), t)\Delta t \tag{2}$$

where Δt is the time step, S_n is a state variable, f_n is the time derivative, or net flux, of S_n at time t and is a function of $\boldsymbol{S}$ and t, and $\boldsymbol{S}$ is a vector of all relevant state variables for computing f_n, determined by the model structure. The flux f is indicated to be a function of $\boldsymbol{S}$ and t because fluxes generally depend on both (for example, net flux to the unsaturated zone depends on precipitation, which is a function of time, and the amount of water in the unsaturated zone, which is a state variable). $O(\Delta t)$ indicates that the truncation error - or the difference between this numerical approximation and an exact solution to the differential equation - is proportional to the size of the time step to the first power. This can be shown by equating the right hand side of Eq. 2 to a Taylor series expansion of the left hand side of Eq. 2, yielding error terms where the lowest order in Δt is 1. Thus this method is first order accurate. The method is known as fixed step because the time step can only be one value (in this experiment, all fixed steps are daily), and explicit because it calculates the flux using state variables that are already known (Süli and Mayers, 2003).

The **fixed step implicit Euler method** is similar to its explicit counterpart in that it is first order accurate and has a fixed time step. It is different in that the derivative at the end of the time step, rather than the start of the time step, is used to calculate the state variable at the end of the time step. However, the derivative of a state variable is generally a function of the state variables, which are initially unknown for the end of the time step. So, the implicit Euler method must iteratively solve a system composed of equations in the style of Eq. 4, searching for the values of $\boldsymbol{S}(t + \Delta t)$ which yield a derivative that matches $S_n(t)$. In the graphical example of Fig. 2b, the value for y at the end of the time step is first approximated with an explicit Euler calculation, and then the method checks if the derivative there leads back to the value at the beginning of the time step, within a given mass balance error tolerance (the mass balance error tolerance and monitoring methods are not studied by this paper). If not, this process is repeated with a new guess for $S_n(t + \Delta t)$ until the derivative at the end of the time step leads back to the value of y at the beginning of the time step, or graphically until the dotted orange line leads back to the value of y
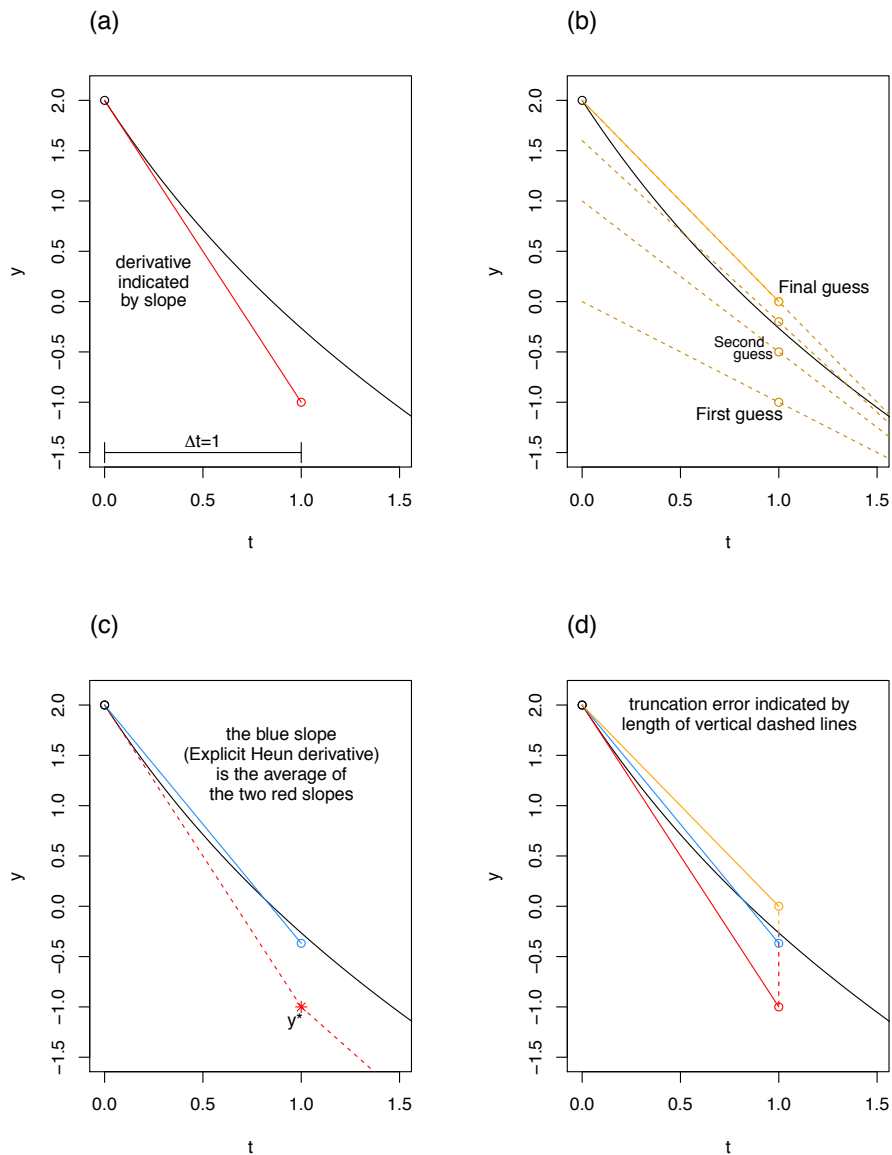
Figure 2. Examples illustrating a single step of explicit Euler (a), implicit Euler (b), explicit Heun (c), and a comparison of all three (d). A single differential equation rather than a system is solved, given by $y' = -y(x(t)) - x(t) - 1$, $y(x(0)) = 2$, where $x(t) = t$ for simplicity. The black curve is the exact solution to this linear differential equation. Each approximation method uses a time step of 1 in these examples.

at $t=0$. Symbolically the method is given by

$$S_n(t + \Delta t) = S_n(t) + f_n(\boldsymbol{S}(t + \Delta t), t + \Delta t)\Delta t + O(\Delta t) \tag{3}$$

$$S_n(t + \Delta t) \approx S_n(t) + f_n(\boldsymbol{S}(t + \Delta t), t + \Delta t)\Delta t \tag{4}$$

where f_n is now the derivative of S_n at time $t + \Delta t$. The iterations required by fixed step implicit methods can make them more computationally expensive than their explicit counterparts. However, a property of implicit methods is that they are unconditionally stable Jameson and Turkel (1981) and can in some situations be more accurate than explicit methods.

The **fixed step semi implicit Euler method** works by first performing an explicit Euler calculation, followed by a single correction in the style of implicit Euler. Visually, this would yield a result at the "second guess" value in Fig. 2b, and mathematically this is achieved by a single iteration of Eq. 4 after the initial estimate of $S_n(t + \Delta t)$. This method then theoretically has a computational expense usually somewhere between the fixed step implicit and explicit Euler methods and is first order accurate. It is typically able to constrain instabilities (Kavetski et al., 2002).

The **fixed step explicit Heun method** works by first explicitly calculating the derivative at the start of a time step (in exactly the same style as fixed step explicit Euler), then explicitly calculating the derivative at the end of the time step using the initial explicit Euler prediction, averaging the derivatives from the start and end of the time step, and then using this corrected average derivative in order to make a final prediction of the state variable value at the end of the time step. In Fig. 2c, the blue slope, or the explicit Heun derivative, is simply the average of both dotted red slopes, where the dotted red slope on the right is calculated using the initial explicit Euler result y^* at $t{=}1$ and the dotted red slope on the left is the same slope as in Fig. 2a. Symbolically,

$$S_n(t + \Delta t) = S_n(t) + \frac{1}{2}[f(\boldsymbol{S}(t), t) + f(\boldsymbol{S}^*(t + \Delta t), t + \Delta t)]\Delta t + O(\Delta t^2) \tag{5}$$

$$S_n(t + \Delta t) \approx S_n(t) + \frac{1}{2}[f(\boldsymbol{S}(t), t) + f(\boldsymbol{S}^*(t + \Delta t), t + \Delta t)]\Delta t \tag{6}$$

where $\boldsymbol{S}^*(t + \Delta t)$ is the initial explicit Euler prediction of the state variables. This is a second order method; its truncation error is proportional to Δt^2, meaning that as Δt approaches 0, the truncation error approaches 0 faster than it would with first order numerical methods.

The final tested fixed step method is the **fixed step implicit Heun method**, also second order. This method is commonly known as the implicit trapezoidal or trapezium method (Süli and Mayers, 2003). It is called the implicit Heun method in this paper due to the fact that it is an implicit analog of the explicit Heun method and for consistency with Clark and Kavetski (2010). The two Heun methods are similar in that they both use the explicit Euler prediction at the beginning of the time step, but the derivative is calculated implicitly in the implicit Heun method, incorporating both the known value of the state variables at time t and the initially unknown value of the state variables at time $t + \Delta t$. Symbolically,

$$S_n(t + \Delta t) = S_n(t) + \frac{1}{2}[f_n(\boldsymbol{S}(t), t) + f_n(\boldsymbol{S}(t + \Delta t), t + \Delta t)]\Delta t + O(\Delta t^2) \tag{7}$$

$$S_n(t + \Delta t) \approx S_n(t) + \frac{1}{2}[f_n(\boldsymbol{S}(t), t) + f_n(\boldsymbol{S}(t + \Delta t), t + \Delta t)]\Delta t \tag{8}$$

where Eq. 8 must be solved iteratively. Graphically, this will appear somewhat similar in method to Fig. 2b.

While so far only fixed step methods have been discussed, this paper also tests adaptive methods, which are able to adapt their step sizes based on various criteria. Adaptive methods can reduce numerical error, due to the fact that a reduced step size will reduce the truncation error. They can also reduce computational expense, because a small step size is not always necessary for high numerical accuracy (Press and Teukolsky, 1992). This paper examines three adaptive methods: the **adaptive semi implicit Euler method**, the **adaptive explicit Heun method**, and the **adaptive implicit Heun method**, all adaptive analogs of their fixed step counterparts. The adaptive semi implicit Euler method adapts its step size by calculating the error e, or the difference between its explicit Euler component and its final value. Specifically, after computing $S_n(t+\Delta t)$, it checks whether

$$e - \tau_r S_n(t+\Delta t) - \tau_a < 0 \tag{9}$$

where τ_r is the relative error tolerance, defaulting in FUSE to a value of 0.01, and τ_a is the absolute error tolerance, defaulting to a value of 0.01 mm. Each state variable must pass this check. If the step size is accepted, then $S_n(t+\Delta t)$ is taken as the state variable for the end of the time step, and the step size is adjusted according to Appendex B in Clark and Kavetski (2010). If the step size is rejected, then the comparison is attempted after the model runs with a smaller step size, also detailed in Appendix B of Clark and Kavetski (2010). The adaptive implicit and explicit Heun methods make the same comparison, although they do calculate e based on differences between a first order component and the second order prediction at the end of the time step, rather than two first order predictions as in the case of the adaptive semi implicit Euler method (Clark and Kavetski, 2010). In all of these adaptive methods, various pre-calculated components are compared, so additional calculations are not needed for the sake of error control. Therefore these are **embedded error control** methods (Press and Teukolsky, 1992).

3 Methods

This study aims to examine the relationship between precipitation extremeness and numerical error for a variety of numerical methods implemented in hydrological models. This section describes the conditions under which models are run, the methods by which numerical error and computational expense are assessed, and the conditions of the literature review by which the approximate magnitudes of errors resulting from popular modeling codes are assessed.

3.1 Modular Modeling Frameworks (MMFs) and the Framework for Understanding Strucutral Errors (FUSE)

Modular modeling frameworks are fairly new tools in hydrology. These are pieces of code or software that are able to rapidly create hydrological models with various processes or structures included, varying numerical schemes (with some MMFs), and other options related to model setup or input. This allows for the controlled comparison of models, enabling studies on model structure, model uncertainty, and a wealth of other topics. In this case, an MMF offers the opportunity to study the change in numerical error resulting from the instantaneous (unrouted) discharge hydrograph as the dimensions of model structure, numerical method, parameters, initial conditions, and forcing data are systematically altered. This lets us test for generality

8

concerning the relation between numerical error and precipitation extremeness over a large dimensional space, as well as account for interactions among these dimensions.

The Framework for Understanding Structural Errors (FUSE) was selected for this experiment. To our best knowledge, this is the only MMF that allows for implicit and explicit numerical methods of higher order than 1. Note that the update to FUSE allowing for higher order numerical techniques was introduced by Clark and Kavetski (2010), after FUSE's initial development (Clark et al., 2008).

Five different FUSE members, or five different model structures, are used in generating all hydrographs. Different combinations of state variables and fluxes are used in each. These include FUSE 070, 563, 550, 092, and 330, which are taken directly from Clark and Kavetski (2010) and are from a list of "models broadly representative of the wide spectrum of conceptual hydrological models used in research and practice." All of these models are lumped rather than distributed. The FUSE snow module is always off in this experiment.

The model runs are generated across twenty parameter sets. This makes the results more generalizable, simulating hydrograph behavior in different physical settings. The parameter sets are generated via a Latin Hypercube (LH) method, ensuring they cover as large a region of the parameter space as possible. The LH technique is used on the total set of parameters used across all models; each model, having a different structure, does not require the full set of LH optimized parameters. In order to determine whether a sufficient number of parameter sets was used, hydrographs resulting from both twenty and eighty parameter sets are separately generated and compared.

3.2 Synthetic Forcing Data

In order to determine the change in model numerical error as precipitation varies from mild to extreme, a clearer idea of what constitutes extreme precipitation is required. Here, 'extreme' precipitation is some combination of intense and long-lasting. In order to test the effect of both changing precipitation intensity and duration on numerical error, precipitation datasets are synthetically generated along the intensity, duration, frequency (IDF) curves shown in Fig. 3. All points in Fig. 3 are World Meteorological Organization world records (Organization, 1994) in total precipitation for a given duration. We generate six IDF curves, representing interpolations between the world record events scaled by factors of 0.01, 0.025, 0.075, 0.25, 1, and 1.2. These factors logarithmically span a large range of rainfall intensities, ranging from mild to larger than historically observed. The inclusion of precipitation events that are more intense than have been historically observed is due to the projected increase in wet day precipitation intensity under emission scenario SRES A2; for many regions, this exceeds 20 % when comparing the simulated periods 2081-2100 and 1980-1999 (Seneviratne et al., 2017). Synthetic precipitation data were generated along these interpolations for three different event durations: 5, 10, and 20 days. These are selected with the motivation that flood modeling is often done for precipitation events at these time scales (Jasper et al., 2002; Weerts and El Serafy, 2006). Thus, in total, 18 synthetic precipitation datasets were generated. The shape of the rainfall signals used in the majority of the analysis is flat, such that each day has the same precipitation intensity for a given event, though the effect of using precipitation data based on real events is briefly discussed. Total accumulated precipitation per event ranges between 40.1 mm (for the 5 day event with
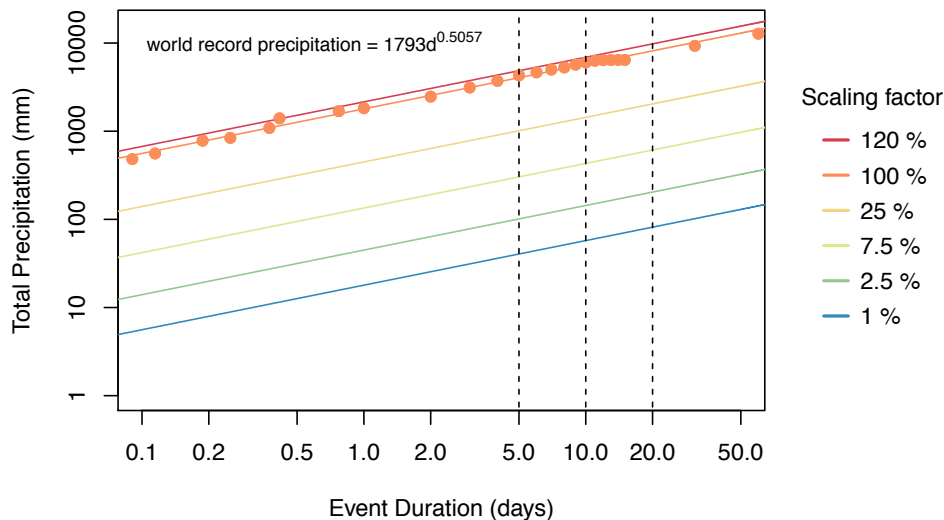
Figure 3. Synthetic IDF (intensity, duration, frequency) curves along which synthetic forcing data were generated, where the vertical black lines show the precipitation event durations used. The IDF curves are interpolated between world record events for given durations, scaled by 6 different factors, ranging from 0.01 to 1.2, indicated by color. The precipitation world records are World Meteorological Organization data (Organization, 1994) and are indicated by orange points. The power function giving the best fit for the world records is shown.

a daily intensity of 1 % of the average intensity of the interpolated 5 day historical maximum) and 9787.2 mm (for the 20 day event with a daily intensity of 120 % of the average intensity of the interpolated 20 day historical maximum).

225   It is also important to determine to what extent results depend on initial conditions. To this end, the forcing data are preceded by a 500 day spin up period with one of three constant precipitation intensities: 2.5 mm d^{-1}, 5 mm d^{-1}, or 10 mm d^{-1}. Each spin up intensity is used with each of the above described precipitation events. The goal of using these three intensities is not necessarily to simulate initial conditions from three distinct climatologies, but rather to establish a broad range of initial conditions in terms of percent of maximum storage. Because we incorporate a broad variety of model structures and parameter

230   sets, the total storage in the model at end of the spin up period spans a variety of values, ranging from nearly empty (less than 1 % full) to approximately 70 % full. From least to most intense spin up periods, median storages expressed as a percent of maximum storages at the end of the spin up period are 21 %, 25 %, and 28 %, respectively. Comparing results from three different spin up precipitation intensities allows for a systematic method by which to investigate initial conditions and their effect on the relative numerical errors of various numerical methods.

235   The synthetic forcing datasets are completed with potential evapotranspiration and temperature data. For all forcing datasets, constant potential ET values of 2 mm d^{-1} and temperatures of 10 degrees Celsius are used. In most cases, the ET flux is small relative to the precipitation flux.

3.3 Method for Determination of Numerical Error

For a fixed model structure (out of 5), initial condition (out of 3 levels), parameter set (out of 20), and forcing dataset (out of 18), there are nine instantaneous discharge hydrograph runs, corresponding to the eight tested numerical methods and the benchmark method. The **benchmark method** is the near-exact solution against which all other methods are compared. It is generated with the most sophisticated numerical technique available in FUSE, adaptive implicit Heun, with error tolerances 1000 times smaller than the defaults $\tau_r = 10^{-5}, \tau_a = 10^{-5}$mm for the benchmark). A desirable property of the benchmark is that it yields approximate solutions that are much closer to the exact solutions than those resulting from any other method. To ensure this, the hydrograph resulting from the benchmark method, but with ten times coarser error tolerances, was compared to the discharge hydrographs from all other methods. It was found that the coarser benchmark had lower error (when compared to the real benchmark) than any other method more than 99 % of the time.

The error for any tested run is determined by calculating the root mean square error (RMSE, $\mathrm{mm\ d^{-1}}$) and normalized root mean square error (NRMSE, percent) with respect to the benchmark run. The RMSE is given by

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{q}_t - q_t)^2} \tag{10}$$

where T is the total time evaluated in days, $\hat{q}_t$ is the near-exact discharge resulting from the benchmark method at time t, and q_t is the discharge resulting from the numerical method being assessed at time t. The NRMSE is simply the RMSE normalized by the mean discharge of the associated benchmark run, then multiplied by 100 to become a percent. The time T on which the error is evaluated is seven days for five day precipitation events, fourteen days for ten day precipitation events, and twenty eight days for twenty day precipitation events, always beginning on the first day of rainfall after the spin up period. This is a somewhat arbitrary albeit valid period on which to evaluate hydrograph performance; it captures all or part of the rising limb, the crest segment, and the falling limb. Figure 4 shows example discharge hydrographs resulting from the fixed step explicit Heun method and the benchmark method. The total number of times the system of differential equations composing a model must be solved, or the number of system flux evaluations, is also recorded and represents the computational expense of generating the hydrograph.

RMSE and NRMSE are both included in this experiment because they show error in two different yet valid ways: in the original units of the discharge and as a relative measure of the near-exact discharge, respectively. RMSE is useful in that its units $(\mathrm{mm\ d^{-1}})$ are easy to interpret, directly showing the numerical error in the daily discharge, while the benefit of NRMSE is that it provides context for the error. These definitions of error are purely numerical and therefore differ from *fidelity* as described in Clark and Kavetski (2010), which further incorporates real discharge observations.

3.4 Literature Review

A small literature review determining available numerical strategies was conducted on twelve hydrological modeling codes. This included seven off-the-shelf models, which are the objects of study in Addor and Melsen (2019), and five popular MMFs. The reviewed modeling codes are shown in Appendix A. Information is gathered on whether recently updated versions of
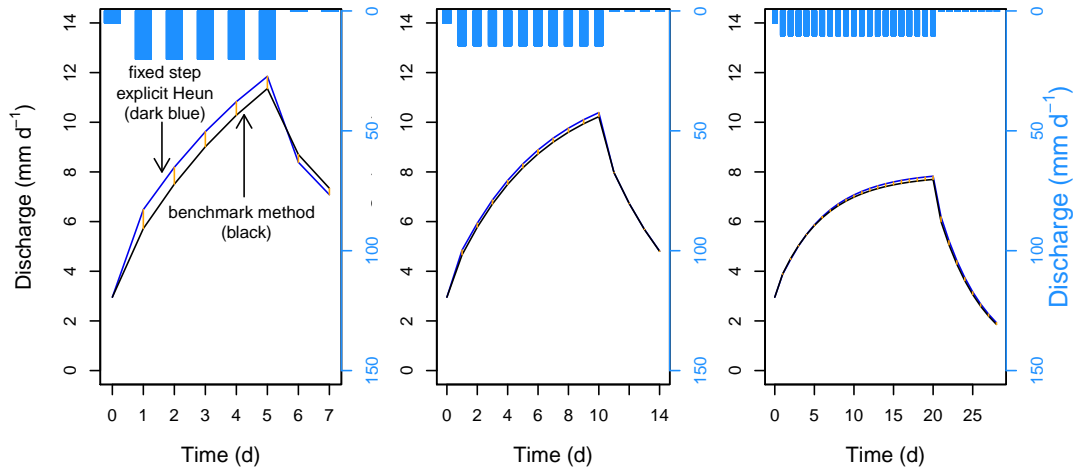
Figure 4. Example hyetographs and discharge hydrographs, including results from the fixed step explicit Heun method and the benchmark method for 5, 10, and 20 day precipitation events. The vertical orange lines are the differences between the two hydrographs and represent numerical error. From shortest to longest duration, the RMSEs are $0.55 \, \mathrm{mm \, d^{-1}}$, $0.14 \, \mathrm{mm \, d^{-1}}$, and $0.10 \, \mathrm{mm \, d^{-1}}$. The precipitation forcing data come from the IDF curve scaled at 2.5 % of the historical maxima interpolation. These hydrographs all result from the same structure (FUSE 092), the same parameter set, and the same spin up precipitation intensity ($5 \, \mathrm{mm \, d^{-1}}$). The last day of the spin up period is shown.

each modeling code are capable of sequential or simultaneous solving, available orders of numerical solver, implicit or explicit (or other) nature of the numerical solver, and whether or not adaptive substepping is available. When solvers were found to have the option of simultaneous solving, all available numerical options were recorded. When solvers were found to only have sequential options - allowing for different numerical options per individual flux - the modeling code is considered to be restricted to the least sophisticated method specified among the fluxes encountered (specifically, we consider explicit methods to be less sophisticated than implicit methods, fixed step less sophisticated than adaptive, and lower order less sophisticated than higher order). This reflects our assumption that a numerically erroneous component of the model will produce errors which propagate to other components of the model; in other words, the numerical accuracy is only as good as the most erroneous part of a model. This review has the objective of approximately placing other models or MMFs in the context of this experiment, which sheds light on what kinds of numerical errors might be expected as a result of using popular modeling codes.

4 Results

4.1 Numerical error and precipitation intensity for 5 day events

Figure 5 details the evolution of numerical errors for the tested numerical techniques for five day events of increasing precipitation intensity, for twenty parameter sets, all model structures, and all initial conditions. It is immediately apparent that for all methods, median RMSE increases with increasing precipitation intensity, and second order adaptive methods usually
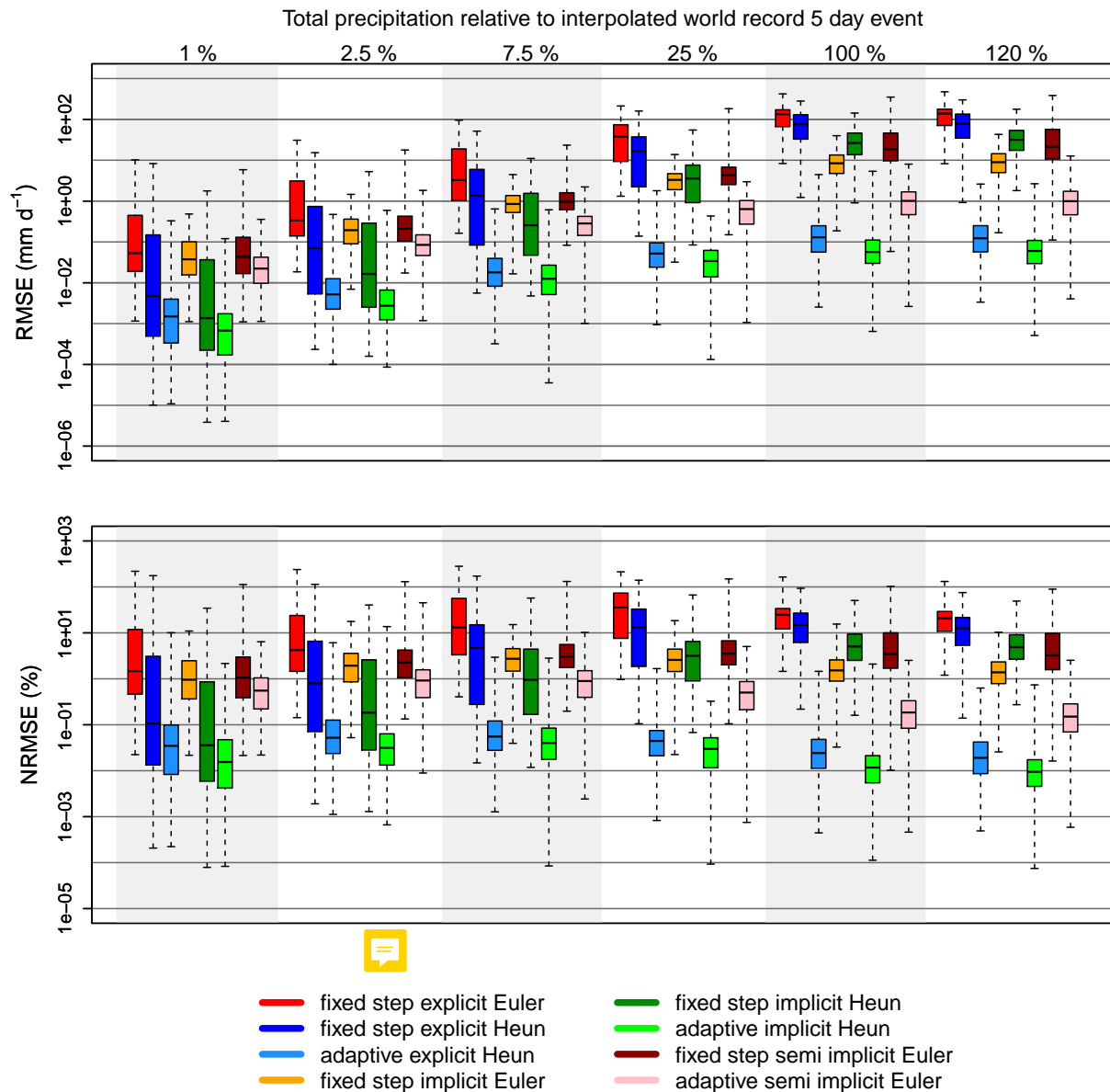
12

Figure 5. Box plots showing the errors (RMSE and NRMSE) associated with each numerical technique for 5 day precipitation events of increasing intensity, where the intensity is represented as a percentage of the interpolated world record 5 day event intensity on the top of the figure. The solid black line in each box is the median error, the boxes extend to the 25 % and 75 % quantiles, and the whiskers extend to extreme values. All adaptive methods use the FUSE default error tolerances. Heun methods are second order, whereas Euler methods are first order. The model runs represented in this figure are generated across 20 parameter sets, 5 model structures, and 3 spin up conditions.

outperform fixed step methods (in some cases by a few orders of magnitude). The following two paragraphs explore trends in RMSE, shown in the top panel of Fig. 5, from left to right.

For lower precipitation intensities, adaptive Heun methods usually outperform the other methods. The adaptive semi implicit Euler method yielded more error than the other adaptive methods, putting this method on par with the second order fixed step methods. The relatively low performance of adaptive semi implicit Euler is likely due to the fact that the Heun methods adapt their step sizes based on comparisons between first and second order results, while the semi implicit method adapts its step size based on a comparison between two first order results. Here, all adaptive methods use the same default error tolerances, where the effects of changing these will be discussed later. Among fixed step methods, for lower precipitation intensities, first order methods (the Euler methods) are on average outperformed by second order methods (the Heun methods). Apparently the extra flux evaluations required by second order methods are able to reduce error in these low intensity cases.

As precipitation intensity increases, it is clear that adaptive methods outperform fixed step methods, where the adaptive semi implicit Euler method begins to yield lower error than fixed step methods. For fixed step methods, with higher precipitation intensity, it is no longer the case that second order methods outperform first order methods. Instead, implicit or semi implicit methods outperform explicit methods; the fixed step explicit methods yield median errors above 70 mm d^{-1} for the two highest precipitation intensities. This indicates that instabilities contribute more significantly to numerical error as precipitation intensity increases, as implicit methods are unconditionally stable (Jameson and Turkel, 1981) and the semi implicit method constrains instabilities more readily than the explicit methods (Kavetski et al., 2002). Furthermore, because fixed step explicit Heun clearly outperforms explicit Euler for low precipitation intensities but not for high precipitation intensities, and because fixed step implicit Heun and implicit Euler have this same property, we can conclude that higher order truncation error is more sensitive to increasing precipitation intensity than first order truncation error at the daily time step. This suggests that simply selecting an arbitrarily high order fixed step method will not always constrain error - and in some cases might make the error worse. Evidently, adaptive methods are most suitable for constraining numerical errors.

In the bottom panel of Fig. 5, the NRMSE is shown. All of the above identified trends can be found here as well, except that median error always increases with increasing precipitation intensity. With many (especially adaptive) methods, NRMSE decreases with increasing precipitation intensity. This is due to the fact that RMSE might usually increase, but not as fast as the discharge by which the RMSE is normalized. When this analysis was repeated with Kling-Gupta Efficiency (Gupta et al., 2009), Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970), and normalized error in maximum discharge, similar trends in performance and precipitation intensity were obtained when compared to NRMSE.

4.2 Robustness of Results

The previous subsection makes a variety of claims relating numerical error and method choice. These are reliant on a ranking of numerical methods, based on numerical error, that is different for low and high precipitation intensity events. For low intensity events, we state that the second order adaptive methods occur as the group with the lowest error, the adaptive semi implicit Euler method and the fixed step second order methods occur as the group with the second lowest error, and the fixed step first order methods occur as the group with the largest error. For high intensity events, we state that the adaptive methods occur

14

as the group with the lowest error, fixed step implicit or semi implicit methods occur as the group with the second lowest error, and fixed step explicit methods occur as the group with the highest error. While these groupings are clearly viable via median errors in RMSE and NRMSE, it could be that various model dimensions can sometimes interact in such a way that these groupings are not observed. In this subsection, we determine how often these groupings are observed in the above ranked orders over the chosen modeling dimensions. In short, we find that the above described rankings are robust over the majority of the tested dimensional space.

When spin up precipitation intensity, forcing dataset, model structure, and parameter set are all fixed to a single choice, there are 8 model runs which have their error calculated, corresponding to the eight tested numerical methods. For each individual set of 8 hydrographs, for the most and least intense 5 day precipitation intensities, methods were given a rank of 1 to 8, based on RMSE, where 1 indicates the lowest RMSE and 8 indicates the highest. Any ranking with ties was discarded, which occurred about 20 % of the time for the lowest intensity and did not occur for the highest intensity. The numerical techniques were sorted into one of the three ranked groups based on their rank, where for the lowest intensity, the two lowest error methods occupy ranked group 1, the next three methods occupy ranked group 2, and the three most erroneous methods occupy ranked group 3, and for the highest intensity, the three lowest error methods occupy ranked group 1, the next three methods occupy ranked group 2, and the two most erroneous methods occupy ranked group 3. Then, the composition of each ranked group is reported. When a ranked group is mostly composed of a single selection of methods, these methods can be said to have the rank corresponding to the ranked group over most of the tested dimensional space. This is indicated by the predominance of a single color for a given ranked group in Fig. 6.

As a concrete example, consider the ranking wherein adaptive second order methods generally outperform other methods for the least intense 5 day precipitation events. This is reliant upon the adaptive Heun methods occurring often in the ranked group with the lowest error (ranked group 1), given the least intense 5 day precipitation data. To determine how often this is the case, each numerical method in each set of 8 runs was given a ranking based on RMSE, where 1 is low error and 8 is high error, and rankings with ties were discarded. Whenever the second order adaptive methods had either rank 1 or 2 in an individual set of 8 runs, they were placed into ranked group 1. As can be seen in Fig. 6, on average, adaptive explicit Heun and adaptive implicit Heun occur in ranked group 1 at a frequency of 74 % across the tested dimensional space for the lowest precipitation intensity for 5 day events, meaning it is usually the case that these two methods as a group have the lowest error for the lowest precipitation intensity for 5 day events. This indicates that the claim "adaptive second order methods outperform other methods for the least intense 5 day precipitation events" is valid over the majority of the dimensional space tested in Fig. 6 (i.e., it is valid over a variety of choices in model structure, parameter set choice, and initial conditions). Note that there are different rankings for the lowest and highest precipitation intensities, as some numerical methods react very differently to increasing precipitation intensity (described in the previous subsection).

Because the frequency with which methods occurred in their desired groups is generally high, we can be confident that the trends in relative method performance in the previous subsection are usually true over the tested model structures, parameter sets, and initial conditions. The fraction of each ranked group composed of its dominant methods always changed by less than 2 % as a result of changing the initial conditions (see Fig. 6). We can conclude that the rankings are relatively insensitive to initial
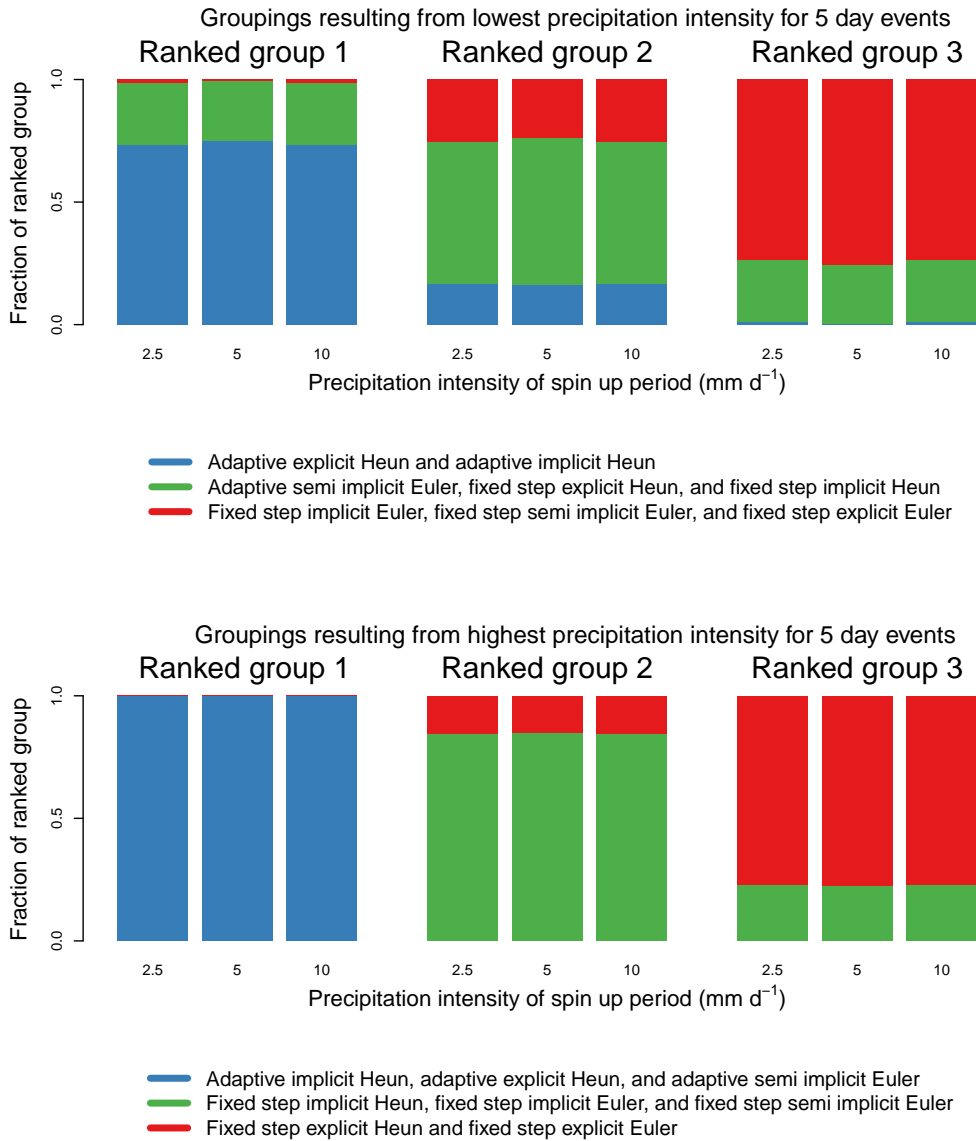
Figure 6. The robustness of rankings of numerical techniques in RMSE for high and low precipitation intensity events, for three different initial conditions. A ranking is more robust when a selection of numerical methods is prevalent in a given ranked group, indicated by the predominance of a single color in a ranked group. All adaptive methods here use FUSE default error tolerances. These results incorporate hydrographs generated across 5 model structures and 20 parameter sets. Different rankings are observed for high and low precipitation intensities.

condition. We determined that the ranking is robust over different model structures as well; the desired ranking was usually observed a high percentage of the time when individual model structures were assessed. However, we found that the model structures containing interflow process representations - namely, FUSE 536, FUSE 550, and FUSE 330 - were less likely than structures without interflow to produce the dominant ranking of Fig. 6, for both the lowest and highest precipitation intensity for 5 day events. The most common deviations from the most common ranking for the lowest precipitation intensity occur due to fixed step implicit Heun outperforming adaptive explicit Heun or adaptive semi implicit Euler performing approximately as well as first order fixed step methods. In order to determine if the rankings were sensitive to number of parameter sets used, hydrographs were further generated over 80 parameter sets, where the spin up precipitation intensity was set to 2.5 mm d^{-1}, with no change in the other dimensions. Then, the rankings were compared between the two cases where 80 or 20 parameter sets were used and the spin up precipitation intensity was 2.5 mm d^{-1}. Because the frequency with which the required ranking is observed changed very little between the two cases - with a largest change for any grouping of 4.9 % - it is evident that the ranking is fairly insensitive to the number of parameter sets used. This indicates that enough parameter sets were used to establish generality over the parameter space. The hydrographs generated over 80 parameter sets were not used beyond investigating ranking robustness.

The same procedure was used to evaluate rankings in computational expense, using number of system flux evaluations rather than RMSE, because further analysis involving computational expense should also be robust over the tested model dimensional space. Five distinct ranked groups were discovered. From lowest to highest expense, these groups include:

1. fixed step explicit Euler;

2. fixed step explicit Heun;

3. adaptive explicit Heun and fixed step semi implicit Euler;

4. fixed step implicit Euler, fixed step implicit Heun, and adaptive semi implicit Euler;

5. adaptive implicit Heun.

This ranking appeared to be common for both low and high precipitation intensity cases. In the style of Fig. 6, and over the same model dimensions, the dominant selection of method(s) never occurred in any ranked group less than 73 % of the time. This indicates that these rankings are robust over a variety of choices in initial condition, parameters, and model structure. Changing either number of parameter sets or initial conditions also always yielded a change in frequency of less than 12 %, demonstrating the insensitivity of computational expense ranking to initial conditions or number of parameter sets used. Based on the results of this subsection, we conclude that the impact of model structure, different parameter sets, and initial conditions on the ranking of numerical methods based on numerical error or computational expense is limited.

4.3 Impact of event duration on error

It was found that longer duration events have lower associated RMSE, where the median RMSE across all methods decreases with increasing duration. This was analyzed across 20 parameter sets, all 5 model structures, and for all 3 spin up precipitation

385 intensities. The average reduction in median RMSE across all methods between 5 and 20 day events was 74 % for the least
intense scaling factor (1 % of interpolated world record average intensity) and 54 % for the most intense scaling factor (120 %
of interpolated world record average intensity), where the median RMSE of all individual methods decreased monotonically
with increasing precipitation duration. This is, to some extent, a mathematical artifact; because the rainfall signal is flat, a larger
portion of a longer time series is closer to an equilibrium discharge. This reduces the average error for the hydrograph. In order
390 to test the extent to which RMSE depends on the geometry of the rainfall signal, the experiment was repeated with new forcing
data based on CAMELS precipitation data observed during hurricane Katrina (Newman et al., 2015). Original forcing data were
scaled such that the total precipitation per event is consistent with the rest of the experiment but relative daily precipitation
intensities reflect real data. Under these conditions, the same dominant compositions of ranked groups are observed (and in fact
become slightly more robust), and the trend of increasing median RMSE with increasing precipitation intensity is still clear.
395 The decrease of error with increasing duration is still present though less pronounced with variable rainfall signals. In this case,
the average reduction in median RMSE across all methods between 5 and 20 day events was 69 % for the least intense scaling
factor (1 % of interpolated world record total precipitation) and 35 % for the most intense scaling factor (120 % of interpolated
world record total precipitation). In either the case of the flat or variable rainfall signal, increasing the duration along an IDF
curve necessarily means reducing the daily intensity of the event. Therefore, the reduction of numerical error due to increasing
400 duration might simply be due to the lower daily intensity per event of longer events. Nonetheless, longer duration events along
an IDF curve do represent more total rainfall per event, and it appears as if models yield less error for longer duration events
along an IDF curve (not shown).

4.4 Impact of rainfall signal geometry on error

Magnitudes of median errors are generally larger with the variable intensity precipitation data but tend to be on the same order
405 of magnitude as in the cases with no variance; because a larger variance in the rainfall signal can increase numerical errors,
we can expect that numerical errors in practice might be somewhat larger than reported in the previous subsection. Numerical
errors resulting from the 5 day forcing data based on hurricane Katrina can be found in Fig. 7.

It is also interesting to note that precipitation intensities greater than the 1 day world record event were contained within
the variable intensity 5 day forcing data. Because these data generally yield larger errors than with the flat rainfall signals,
410 we can conclude that shorter duration events incorporating larger intensities can produce larger numerical errors. This further
establishes the relationship between numerical error and precipitation intensity and especially underscores the importance of
numerical technique choice for intense, short duration events.

4.5 Computational efficiency

So far, this study makes assessments of numerical errors associated with various numerical techniques. As a result, and with
415 default error tolerances, it seems as if adaptive second order methods tend to yield the lowest numerical error. Low error is
desirable, but computational expense can also be a decisive factor when selecting a numerical method. Figure 8 shows the
computational efficiency of the tested numerical techniques. More efficient methods will produce a relatively low median
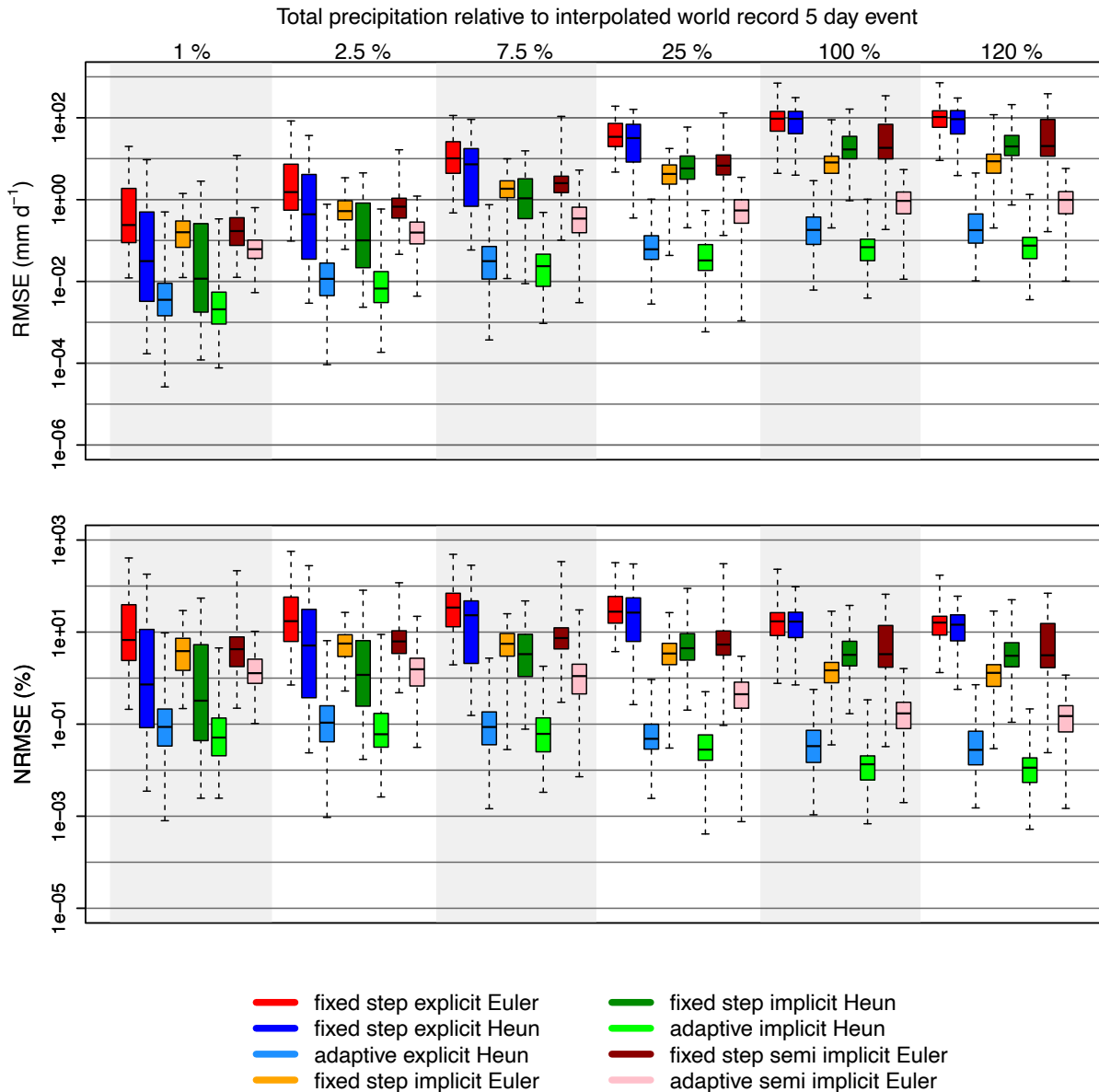
18

Figure 7. Box plots showing the errors (RMSE and NRMSE) associated with each numerical technique for 5 day precipitation events of variable intensity with increasing total precipitation, where the total precipitation is represented as a percentage of the interpolated world record 5 day event total precipitation on the top of the figure. The solid black line in each box is the median error, the boxes extend to the 25 % and 75 % quantiles, and the whiskers extend to extreme values. All adaptive methods use the FUSE default error tolerances. The model runs represented in this figure are generated across 20 parameter sets, 5 model structures, and 3 spin up conditions. Note that errors are generally larger than in the case where the precipitation signals have no variance per event.
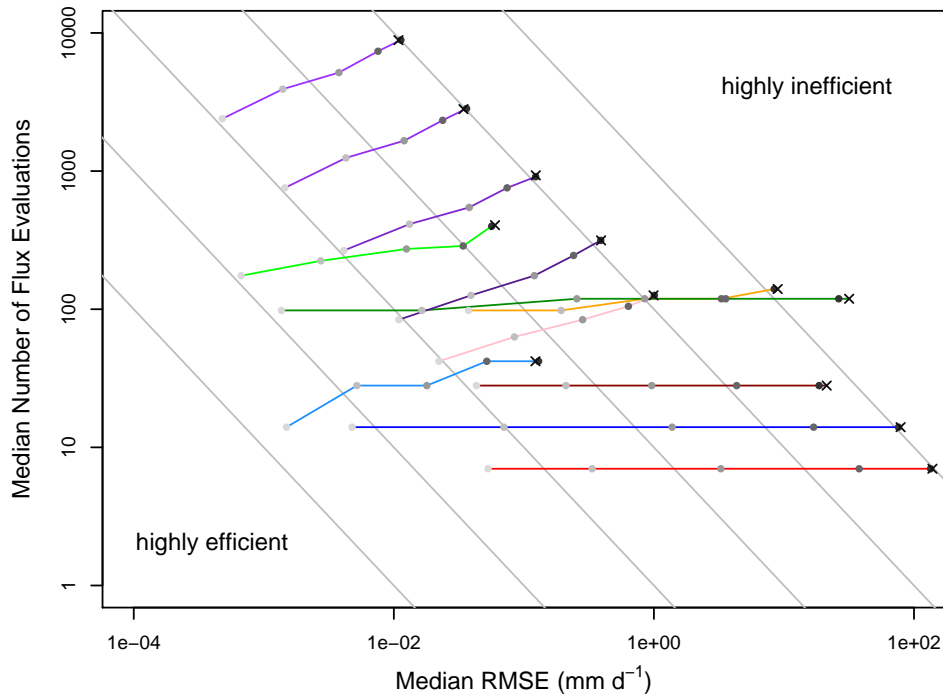
Figure 8. Relationship between computational efficiency and the numerical error, in the style of a Pareto front, where methods are more efficient if they have a low RMSE and a low number of flux evaluations. All medians are calculated across all 5 model structures, 20 parameter sets, and 3 spin up precipitation intensities. All six precipitation intensities for 5 day events are shown, where the light gray dot represents the lowest intensity event, darker gray dots represent increasingly intense precipitation events, and the dark gray 'x' represents the highest intensity event. Unless otherwise stated, adaptive methods use the FUSE default tolerances. Purple lines depict adaptive semi implicit Euler methods with error tolerances below the FUSE default, where the light purple method has the smallest error tolerance. Gray lines are of the form $y = \frac{10^n}{x}$, where n is an integer, such that moving between adjacent gray lines represents an order of magnitude change in efficiency.

numerical error and will take a relatively small median number of flux evaluations in order to achieve this, i.e. they will occur closer to the leftmost gray line. Mathematically, we can also consider high efficiency to be achieved when the product of 420 number of function calls and numerical error is minimized. We explore multiple error tolerances for the adaptive semi implicit

Euler method. These are examined in order to determine how efficient the adaptive semi implicit method is when it yields approximately as much error as the adaptive Heun methods with default error tolerances.

Because each precipitation intensity is shown for 5 day events, it is clear that the efficiency of each method tends to decrease with increasing precipitation intensity. Arbitrarily small error can be achieved with adaptive methods by selecting smaller error tolerances. The semi implicit method with the smallest error tolerance has the lowest error of any tested numerical method, though this comes at a large computational cost. Adjusting the error tolerances of the adaptive semi implicit method does not seem to change its efficiency. The central feature of Fig. 8 is that the adaptive explicit Heun method emerges as a clear leader in terms of efficiency, significantly outperforming all other methods for any precipitation intensity. This was also established for the longer duration events (not shown).

It is possible to examine individual sets of 8 runs, rather than median errors and expenses, to determine how often the adaptive explicit Heun method has the highest efficiency. We find that the adaptive explicit Heun method is the most efficient method among those tested 85 % of the time, on average over all 18 forcing datasets, 3 initial condition levels, 20 parameter sets, and 5 model structures. When model runs were grouped based on percentage of world record precipitation intensity, it was found that percentage of the model runs for which the adaptive explicit Heun method was most efficient ranged between 80.3 % (when precipitation intensities were at 1 % of historical maxima) and 90.1 % (when precipitation intensities were at 25 % of historical maxima).

4.6 Numerical choices control numerical error more than structural choices

Earlier, numerical error was defined as differences between the exact and approximate solutions to the set of equations composing the model that result from the choice of numerical method used to find an approximate solution. While numerical error ultimately comes from numerical method choice, the same method in different structural contexts might produce different error magnitudes, so model structure might be able to contribute to numerical error.

A basic ANOVA analysis was conducted on NRMSE for all model runs using 20 parameter sets, all 3 spin up precipitation intensities, and with all 18 forcing datasets. P-values for groupings based on numerical method choice and structural choice were calculated. The p-values for groupings based on numerical method were always less than 1e-20, demonstrating that numerical method choice strongly controls NRMSE. The p-values for groupings based on structure were below 0.05 for zero of the six 5 day forcing datasets, two of the six 10 day forcing datasets, and three of the six 20 day forcing datasets. The p-values for numerics groupings were always lower than the p-values for structural groupings. This demonstrates that numerical choices control numerical error more strongly than model structural choices, and structural choices seem to become more important in controlling numerical error for longer duration events.

4.7 Literature review results

Finally, a short literature review was conducted to place the results of this study in context. The surveyed hydrological modeling codes were lumped into three categories, shown in Fig. 10. Presently we discuss numerical errors that might result from each category.
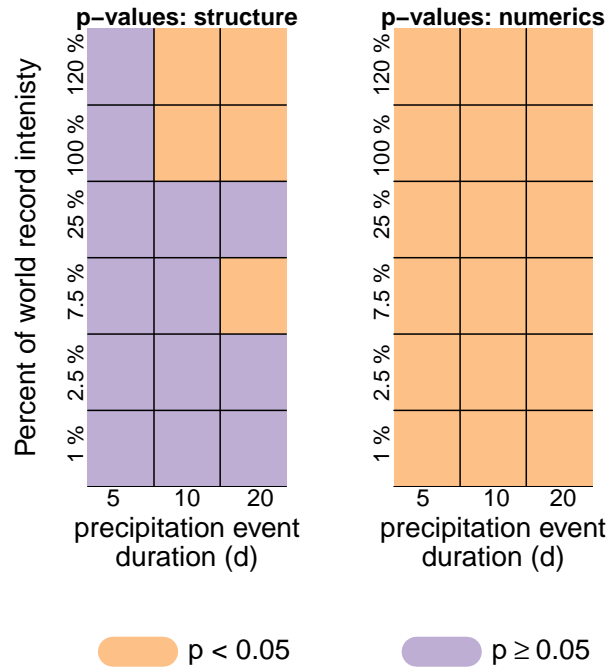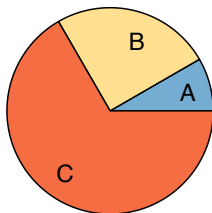
Figure 9. Statistical significance for groupings based on structural choices (left) and numerical choices (right). The p-values are the results of ANOVA tests, which were conducted on NRMSE.

(A) a second order explicit adaptive method can be used;

(B) an adaptive method, but not a higher order explicit method, can be used;

(C) a fixed step method is used.

Figure 10. Grouping the reviewed modeling codes (12 in total) into three categories, based on available numerical techniques. Category C codes are the most common.

Of the twelve reviewed codes, one belongs to category A, three belong to category B, and eight belong to category C; the highest order numerical method found in any code was second order. The codes of categories A and B, both having the option of adaptive substepping, can guarantee any level of numerical accuracy with sufficiently small error tolerances. The difference is that the numerical techniques of category B did not include higher order (second order or higher) adaptive explicit methods and so are not identified as optimally efficient, according to Fig. 8.

The codes of category C have fixed step solvers. These codes can be further subdivided into more categories, based on the ability or inability to specify a fixed time step, based on sequential or simultaneous solving, based on order of numerical

method, or based on implicit or explicit (or other) nature of the numerical methods available. However, the fixed time step either causes an extremely large computational expense when the fixed time step can be arbitrarily set by the user to constrain error (Clark and Kavetski, 2010; Press and Teukolsky, 1992) or relatively large numerical inaccuracies when the time step can not be adjusted by the user (as shown by this experiment, at least for the daily time step).

The most common method of category C is closely analogous to the fixed step explicit Euler method, albeit in sequential rather than simultaneous form. If willing to accept the flat rainfall signal and chosen dimensional space of Fig. 5 as physically possible, and if willing to accept that FUSE can at least approximately mimic other codes, one could tentatively estimate that category C codes could often yield large median numerical errors, from the perspective or RMSE or NRMSE. Specifically, the fixed step explicit Euler method yields a median NRMSE of more than 4 % for 5 day events where the precipitation intensity is 2.5 % of the average intensity for the 5 day historical maximum. This is already fairly large, but median NRMSE for this method can become much larger; by the time precipitation is at 25 % of the world record, numerical NRMSE is over 30 %. Similar trends can be found for longer duration events, where median errors in RMSE and NRMSE for this method increase with increasing intensity. For 10 and 20 day events, the median NRMSE resulting from the explicit Euler method is maximized at 29 % and 23 %, respectively. When judging a numerical technique by its extremes in error rather than its average, the fixed step explicit Euler method performs poorly; in this experiment, extremes easily exceed 100 % NRMSE for any forcing dataset. These results seem to indicate that the most frequently used numerical technique that we encountered in our short survey performs relatively poorly for any precipitation event, where its performance is especially poor for (increasingly common) extreme precipitation events.

5   Discussion

5.1   Implications of the study design and further opportunities

Model dimensional space covered. Though the dimensional space tested in this experiment is reasonably large (incorporating various model structures, parameter sets, initial conditions, forcing data, and numerical techniques), it does not cover all possible behaviors of its dimensions. There are model structures in common use outside of the five tested, especially considering that we do not incorporate processes related to freezing or snow. Additionally, we do not use an energy balance, rather FUSE generates bucket style hydrological models that only solve a mass balance. Because land surface models incorporate processes involving energy, these might have a different (perhaps more complicated) relationship between numerical error and precipitation extremeness.

Further, these results only apply to lumped rainfall runoff models; incorporating distribution in forcing data or model structure could introduce further complexity that is not considered by this experiment. It is important to note that all models generated by FUSE are composed of systems of nonlinear first-order ordinary differential equations; these could have different error characteristics than partial differential equations (such as the Richards equation or shallow wave equations) or higher-order differential equations. Next, even higher order or differently implemented low order numerical methods, more initial conditions, less intense precipitations, or longer or shorter precipitation durations remain untested. Still, the results were shown to be rather

23

robust across the fairly broad tested dimensional space, which indicates that extrapolation of the results to other precipitation durations and intensities might be justified.

All rainfall signals in this experiment are either constant with respect to time or are based on a single real event with variance. As such, we do not account for a broad variety of precipitation regimes in terms of variable intensity, which might introduce complexity that is not accounted for in this experiment. However, similar results were obtained with different rainfall signal geometries, indicating the results might be robust across rainfall geometries. Furthermore, the trends relating numerical error resulting from a numerical method and precipitation extremeness appeared to be rather insensitive to the other tested dimensions, providing robustness to the conclusions.

Because we used a Latin hypercube approach in generating parameter sets and demonstrated that using a larger number of parameter sets did not significantly change the results, we can conclude that we have sufficiently sampled the parameter space. However, in practice, it is likely that most parameters obey a given distribution, which would limit the parameter space which represents physically likely catchments. By sampling broadly rather than using calibrated parameter sets, we do not necessarily represent catchments that are commonly observed. Both approaches have merit: using only calibrated parameter sets might provide a greater assurance that results arise from only physically likely catchments but could bias the results via parameter restriction, while using a broad sampling of the parameter space gives more weight to catchments that might be underrepresented but might not proportionately represent average or common physical settings. Nonetheless, a large variety of parameter combinations is used in Monte Carlo optimization methods regardless of physical realism, which further makes the investigation of numerical error resulting from a broad sampling of parameters worthwhile. While this experiment uses a broad sampling of the parameter space, using only calibrated parameter sets is also a defensible choice - although there is a risk for interaction between numerical error and calibration optimization.

Numerical details. Several numerical details remain untested by this experiment. First, FUSE employs embedded error control, which is more efficient than error control requiring extra calculations by about a factor of 2 (Press and Teukolsky, 1992). While we can assume that embedded error control methods are relatively efficient, unembedded methods remain untested, and it is possible that some hydrological models do not use embedded methods. It is possible that this could alter the relative efficiency of various adaptive methods, albeit probably not to a different order of magnitude. Second, all implicit techniques here use a single method to ensure that mass balance errors are sufficiently small; changing the treatment of mass balance errors could lead to somewhat different median computational expenses for implicit methods. However, the median number of flux evaluations per time step under a variety of mass balance error monitoring strategies for fixed step implicit Euler apparently ranges between approximately 8 and 21 (Clark and Kavetski, 2010), where the median number of flux evaluations for fixed step implicit Euler in 5 day events ranges between 14 and 20 (from lowest to highest precipitation intensities) in this experiment. This suggests that the implicit Euler method studied in this paper would probably not achieve a new order of magnitude of efficiency as shown in Fig. 8, regardless of mass balance error monitoring method. Third, this study does not examine different methods for enforcing solution constraints (e.g., making sure physically impossible storages do not occur as model results). The method for enforcing solution constraints is outlined by Clark and Kavetski (2010), and altering this method could affect numerical error (Shampine et al., 2005). Finally, the adaptive explicit Heun method appears to be the most efficient technique

given the limited space of techniques tested; this may no longer be the case when even higher order adaptive methods are studied. For example, adaptive fourth or fifth order Runge-Kutta methods (Schoups et al., 2010) are untested here. A truly optimal solver might be able to dynamically change its order or other options associated with the solver, as in other fields (Karimov et al., 2017; Rackauckas et al., 2020; Lauritzen et al., 2010; Ullrich et al., 2017). Though this point is technically a subset of issues with total dimensional space sampling, testing even higher order or further modularized numerical techniques would be a straightforward and useful advancement of this work.

Breadth of literature review. The literature review is not an exhaustive survey of all existing modeling codes, rather it is a smaller investigation of some of the modeling codes that we consider to be widely used. Therefore, it should not be interpreted as a quantitative result of how common numerical errors are, but rather a preliminary indication of what might be. Further note that we only suggest the magnitude and prevalence of errors arising from model runs that are not further scrutinized. Accordingly, our results give an indication of numerical errors arising purely from modeling codes, rather than the occurrence of these errors in published work. For example, it has been demonstrated that fixed step explicit methods are likely to produce large instabilities, but these would probably be identified as such and discarded by a majority of model users.

5.2 Are minimized numerical error and computational expense good metrics by which to choose a numerical method?

At a glance, it might seem as if an inexpensive numerical method with minimized numerical error - the adaptive explicit Heun method - would be an optimal choice for modeling. After all, Clark and Kavetski (2010) do show that numerical error can easily be a large source of error in a model. However, a numerical technique that minimizes numerical error might not always minimize total error in an observational context; in reality, numerical error can cancel out other error sources. This is possible, for example, when a model structure is not sufficiently diffuse, and then a first order fixed step numerical choice is overly diffuse. The numerical method introduces numerical diffusion (a numerical error), which interacts with the structural error in such a way that the total error is reduced when evaluated on discharge observations, even when compared to a near-exact solution (Clark and Kavetski, 2010). Numerical diffusion has even been intentionally used to represent physical diffusion for increased accuracy (Thober et al., 2019). Further, during calibration, error due to parameter choice can cancel out numerical error, in the same way model structural error can.

Whether or not these are desirable qualities of a numerical method (i.e., should hydrological models simply provide accurate predictions of observable data, or should they accurately represent their intended structure regardless of physical realism) is an open question, where the answer is context-dependent. On one hand, a more accurate real-world prediction is ultimately more useful for anyone whose livelihood depends on model results, e.g. a municipality which has to decide whether or not to issue evacuation advisories due to flooding risk. On the other hand, having somewhat unknown errors cancel each other out is a classic instance of getting the right answer for the wrong reasons (Kirchner, 2006), which could pose a problem when the aim of a modeling effort is to gain understanding of hydrological processes. Especially because this experiment shows that model structure can impact numerical error, it is possible that a structural development in a model yielding reduced error is interpreted as an improved representation of nature, where in reality the structural development changed the numerical error

such that the total error is reduced. When we want to test multiple processes based hypotheses, it would be ideal if we could be more assured that our results are due to the structural changes we make, rather than conflated with numerical error. With this paper, we do not claim to have identified optimal methods in an observational context. We do submit, however, that we have identified a few numerical techniques that are desirable (or undesirable) for future process based hypothesis testing.

5.3 Recommendation for numerical technique

Of the studied methods, this experiment finds that the adaptive explicit Heun method provides the best economy of low error and computational expense on average and accordingly is a practical choice for numerical method. This is consistent with Clark and Kavetski (2010) and Kavetski and Clark (2010). However, this study differs in that it does not recommend the fixed step implicit Euler method as an optimal solution, due to the fact that this method on average has relatively large numerical error and computational expenses here. The difference in recommendation between the present study and Clark and Kavetski (2010) or Kavetski and Clark (2010) is likely due to two methodological differences. First, Clark and Kavetski (2010) and Kavetski and Clark (2010) use the concept of fidelity to determine how closely a numerical technique approximates an exact solution, which incorporates real discharge observations. The present study uses purely numerical definitions of error, which are simply different methods. Second, the time periods on which errors were evaluated were different, where the present study evaluates error on the time scale of days and Clark and Kavetski (2010) and Kavetski and Clark (2010) evaluate error on the time scale of years. As this study shows, RMSE depends on precipitation intensity and duration. It could be the case that the fixed step implicit Euler method produces small numerical errors for a majority of a long time series when precipitation is mild or absent, which would keep the numerical error small on average. In contrast, a shorter but more intense precipitation time series yields larger average numerical errors. On balance, it is encouraging that the present study and the 2010 numerical daemon papers authored by Clark and Kavetski agree that the fixed step explicit Euler method produces a large amount of error and that the adaptive explicit Heun method is a good choice.

6 Conclusions

In this study, the effects of changing precipitation intensity and duration on the numerical errors and computational expenses resulting from various numerical techniques in the context of lumped hydrological models were examined. Precipitation events varying between 1 % and 120 % of interpolated historical precipitation intensity maxima for various durations were used as forcing data. Model results were generated over a variety of choices in parameters, initial conditions, and model structures for generality. It was found that median root mean square error (RMSE) usually increased with increasing precipitation intensity and decreased with increasing precipitation event duration. Rankings of numerical techniques in terms of RMSE and number of flux evaluations were established for varying precipitation intensities. These were shown to be robust over the majority of choices in parameters, model structures, and initial conditions. Then, the efficiency of each technique was examined by comparing computational expense and numerical error. Of the tested methods, a clear leader in efficiency emerged for all precipitation intensities: the adaptive explicit Heun method. A small literature review was conducted on various hydrological

595 modeling codes, demonstrating that this method and even adaptive methods in general might be uncommon in practice, and that a much more numerically erroneous method might be common. Because process based hydrological knowledge is enhanced by reductions in numerical error and computational expense, and because numerical errors are in many areas likely to increase with the changing climate (via more extreme precipitation), we advocate a more widespread use of highly efficient numerical methods.

7 Data availability

600

The data used in producing this manuscript are available at https://www.hydroshare.org/resource/acf9a56e93bc4863b972e1e8af36dc0c/ (La Follette, 2020).

Appendix A: Reviewed codes

The MMFs reviewed for this study were the Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) (Knoben,
605 2018; Knoben et al., 2019), the Structure for Unifying Multiple Modeling Alternatives (SUMMA) (NCAR, 2017; Clark et al., 2015), Raven (Snowdon, 2010; Craig et al., 2020), the Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008; Clark and Kavetski, 2010), and SUPERFLEX (Fenicia et al., 2011), and the models reviewed for this study included the Variable Infiltration Capacity model (VIC) (University of Washington, 2016; Hamman et al., 2018), the mesoscale Hydrological model (mHM) (Zink and Cuntz, 2013; Samaniego et al., 2010), TOPMODEL (Buytaert, 2018; Metcalfe et al., 2015;
610 Kavetski et al., 2003), the Precipitation Runoff Modelling System (PRMS) (Markstrom et al., 2015), the Génie Rural model à 4 paramètres Journaliers (GR4J) (Perrin et al., 2003; Santos et al., 2018), the Sacramento soil moisture accounting model (Finnerty et al., 1997; Burnash et al., 1973; Koren et al., 2014), and Hydrologiska Byråns Vattenbalansavdelning (HBV) light (Reynolds et al., 2017). The most recent publication displaying detailed numerics for TOPMODEL indicates a fixed step, and accordingly TOPMODEL is included in category C. However, TOPMODEL has been implemented with an adaptive second
615 order explicit method before this (Kavetski et al., 2003). It could instead be included in category A, if the literature review was redefined to not simply assess the most recent numerical method for a code.

27

as comments and suggestions which improved the way in which initial conditions were set. KJ provided necessary information
625 for the literature review (along with PL), as well as reviews throughout the writing process.

References

Addor, N. and Melsen, L.: Legacy, rather than adequacy, drives the selection of hydrological models, Water Resources Research, 55, 378–390, 2019.

Addor, N., Jaun, S., Fundel, F., and Zappa, M.: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios, Hydrology and Earth System Sciences, 15, 2327–2347, 2011.

Burnash, R., Ferral, R., and McGuire, R.: A generalised streamflow simulation system–conceptual modelling for digital computers. Joint Federal and State River Forecast Center, Tech. rep., Sacramento, Technical Report, 1973.

Buytaert, W.: Implementation of the Hydrological Model TOPMODEL in R, https://cran.r-project.org/web/packages/topmodel/topmodel.pdf, 2018.

Clark, M. P. and Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, Water Resources Research, 46, 2010.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, 2008.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., et al.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resources Research, 51, 2498–2514, 2015.

Coxon, G., Freer, J., Wagener, T., Odoni, N., and Clark, M.: Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments, Hydrological Processes, 28, 6135–6150, 2014.

Coxon, G., Freer, J., Lane, R., Dunne, T., Knoben, W. J., Howden, N. J., Quinn, N., Wagener, T., and Woods, R.: DECIPHeR v1: Dynamic fluxEs and ConnectIvity for Predictions of HydRology., Geoscientific Model Development, 12, 2019.

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, W., Jost, G., Lee, K., Mai, J., Serrer, M., Snowdon, A. P., Sgro, N., et al.: Flexible watershed simulation with the Raven hydrological modelling framework, Environmental Modelling & Software, p. 104728, 2020.

Dralle, D. N., Hahm, W. J., Rempe, D. M., Karst, N. J., Thompson, S. E., and Dietrich, W. E.: Quantification of the seasonal hillslope water storage that does not drive streamflow, Hydrological processes, 32, 1978–1992, 2018.

Fenicia, F., Kavetski, D., and Savenije, H. H.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47, 2011.

Finnerty, B. D., Smith, M. B., Seo, D.-J., Koren, V., and Moglen, G. E.: Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs, Journal of Hydrology, 203, 21–38, 1997.

Glowinski, R., Osher, S. J., and Yin, W.: Splitting methods in communication, imaging, science, and engineering, Springer, 2017.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, 2009.

Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., and Mao, Y.: The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility, Geoscientific Model Development (Online), 11, 2018.

Higham, N. J.: Accuracy and stability of numerical algorithms, vol. 80, Siam, 2002.

Huang, X., Swain, D. L., and Hall, A. D.: Future precipitation increase from very high resolution ensemble downscaling of extreme atmospheric river storms in California, Science Advances, 6, 2020.

Jameson, A. and Turkel, E.: Implicit schemes and LU Decompositions, Mathematics of Computation, 37, 385–397, 1981.

Jasper, K., Gurtz, J., and Lang, H.: Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model, Journal of hydrology, 267, 40–52, 2002.

670 Karimov, A. I., Butusov, D. N., and Tutueva, A. V.: Adaptive explicit-implicit switching solver for stiff ODEs, in: 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 440–444, IEEE, 2017.

Kavetski, D. and Clark, M. P.: Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, Water Resources Research, 46, 2010.

Kavetski, D., Binning, P., and Sloan, S. W.: Noniterative time stepping schemes with adaptive truncation error control for the solution of
675 Richards equation, Water Resources Research, 38, 29–1, 2002.

Kavetski, D., Kuczera, G., and Franks, S. W.: Semidistributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC, Water resources research, 39, 2003.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, 2006.

680 Kirchner, J. W.: Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, Water Resources Research, 45, 2009.

Kittel, C. M. M., Nielsen, K., Tøttrup, C., and Bauer-Gottwein, P.: Informing a hydrological model of the Ogooué with multi-mission remote sensing data, Hydrology and Earth System Sciences, 22, 1453–1472, 2018.

Knoben, W.: wknoben/MARRMoT, https://github.com/wknoben/MARRMoT/tree/master/MARRMoT/Functions/Timestepping, 2018.

685 Knoben, W. J., Freer, J. E., Fowler, K. J., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall-Runoff Models Toolbox (MAR-RMoT) v1. 2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, 2019.

Koren, V., Smith, M., and Cui, Z.: Physically-based modifications to the Sacramento Soil Moisture Accounting model. Part A: Modeling the effects of frozen ground on the runoff generation process, Journal of hydrology, 519, 3475–3491, 2014.

690 La Follette, P.: Data for La Follette et al (Numerical Daemons and extreme precipitation), http://www.hydroshare.org/resource/acf9a56e93bc4863b972e1e8af36dc0c, 2020.

Lauritzen, P. H., Jablonowski, C., Taylor, M. A., and Nair, R. D.: Rotated versions of the Jablonowski steady-state and baroclinic wave test cases: A dynamical core intercomparison, Journal of Advances in Modeling Earth Systems, 2, 2010.

Madsen, H., Lawrence, D., Lang, M., Martinkova, M., and Kjeldsen, T.: Review of trend analysis and climate change projections of extreme
695 precipitation and floods in Europe, Journal of Hydrology, 519, 3634–3650, 2014.

Markstrom, S. L., Regan, R. S., Hay, L. E., Viger, R. J., Webb, R. M., Payn, R. A., and LaFontaine, J. H.: PRMS-IV, the precipitation-runoff modeling system, version 4, US Geological Survey Techniques and Methods, 2015.

Meehl, G. A., Arblaster, J. M., and Tebaldi, C.: Understanding future patterns of increased precipitation intensity in climate model simulations, Geophysical Research Letters, 32, 2005.

700 Melsen, L. and Guse, B.: Hydrological drought simulations: How climate and model structure control parameter sensitivity, Water Resources Research, 55, 10,527–10,547, 2019.

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J., Clark, M. P., Uijlenhoet, R., and Teuling, A. J.: Mapping (dis) agreement in hydrologic projections, Hydrology and Earth System Sciences, 22, 1775–1791, 2018.

Metcalfe, P., Beven, K., and Freer, J.: Dynamic TOPMODEL: A new implementation in R and its sensitivity to time and space steps,
705 Environmental Modelling & Software, 72, 155–172, 2015.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

NCAR: NCAR/summa, https://github.com/NCAR/summa/blob/master/docs/input_output/SUMMA_input.md#infile_model_decisions, 2017.

Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J., et al.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209, 2015.

Noh, S. J., Rakovec, O., Weerts, A. H., and Tachikawa, Y.: On noise specification in data assimilation schemes for improved flood forecasting using distributed hydrological models, Journal of hydrology, 519, 2707–2721, 2014.

Organization, W. M.: Guide to hydrological practices, Secretariat of the World Meteorological Organization, 1994.

Paton, F., Maier, H., and Dandy, G.: Relative magnitudes of sources of uncertainty in assessing climate change impacts on water supply security for the southern Adelaide water supply system, Water Resources Research, 49, 1643–1667, 2013.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of hydrology, 279, 275–289, 2003.

Prancevic, J. P. and Kirchner, J. W.: Topographic controls on the extension and retraction of flowing streams, Geophysical Research Letters, 46, 2084–2092, 2019.

Prein, A. F., Rasmussen, R. M., Ikeda, K., Liu, C., Clark, M. P., and Holland, G. J.: The future intensification of hourly precipitation extremes, Nature Climate Change, 7, 48–52, 2017.

Press, W. H. and Teukolsky, S. A.: Adaptive Stepsize Runge-Kutta Integration, Computers in Physics, 6, 188–191, 1992.

Rackauckas, C., Ranocha, H., Thakur, D., et al.: ODE Solvers, https://github.com/SciML/DiffEqDocs.jl/blob/master/docs/src/solvers/ode_solve.md, 2020.

Refsgaard, J. C.: Parameterisation, calibration and validation of distributed hydrological models, Journal of hydrology, 198, 69–97, 1997.

Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resources Research, 32, 2189–2202, 1996.

Rempe, D. M. and Dietrich, W. E.: A bottom-up control on fresh-bedrock topography under landscapes, Proceedings of the National Academy of Sciences, 111, 6576–6581, 2014.

Reynolds, J., Halldin, S., Xu, C.-Y., Seibert, J., and Kauffeldt, A.: Sub-daily runoff predictions using parameters calibrated on the basis of data with a daily temporal resolution, Journal of hydrology, 550, 399–411, 2017.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 2010.

Santos, L., Thirel, G., and Perrin, C.: Continuous state-space representation of a bucket-type rainfall-runoff model: a case study with the GR4 model using state-space GR4 (version 1.0), 2018.

Schoups, G., Vrugt, J., Fenicia, F., and Van de Giesen, N.: Corruption of accuracy and efficiency of Markov chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, Water Resources Research, 46, 2010.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., et al.: Changes in climate extremes and their impacts on the natural physical environment, 2017.

Shampine, L. F., Thompson, S., Kierzenka, J., and Byrne, G.: Non-negative solutions of ODEs, Applied mathematics and computation, 170, 556–569, 2005.

Snowdon, A.: Improved numerical methods for distributed hydrological models, Master's thesis, University of Waterloo, 2010.

745   Süli, E. and Mayers, D. F.: An introduction to numerical analysis, Cambridge university press, 2003.

Thober, S., Cuntz, M., Kelbling, M., Kumar, R., Mai, J., and Samaniego, L.: The multiscale routing model mRM v1. 0: simple river routing at resolutions from 1 to 50 km, Geoscientific Model Development, 12, 2501–2521, 2019.

Trenberth, K. E.: Changes in precipitation with climate change, Climate Research, 47, 123–138, 2011.

Ullrich, P. A., Jablonowski, C., Kent, J., Lauritzen, P. H., Nair, R., Reed, K. A., Zarzycki, C. M., Hall, D. M., Dazlich, D., Heikes, R.,
750   et al.: DCMIP2016: a review of non-hydrostatic dynamical core design and intercomparison of participating models, Geoscientific Model Development, 10, 4477–4509, 2017.

University of Washington, C. H. G.: Frequently Asked Questions, https://vic.readthedocs.io/en/master/FAQ/FAQ/, 2016.

Weerts, A. H. and El Serafy, G. Y.: Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, Water resources research, 42, 2006.

755   Wrede, S., Fenicia, F., Martínez-Carreras, N., Juilleret, J., Hissler, C., Krein, A., Savenije, H. H., Uhlenbrook, S., Kavetski, D., and Pfister, L.: Towards more systematic perceptual model development: a case study using 3 Luxembourgish catchments, Hydrological processes, 29, 2731–2750, 2015.

Zink, M. and Cuntz, M.: mhm.nml · develop · mhm / mhm, https://git.ufz.de/mhm/mhm/blob/develop/mhm.nml, 2013.