Neglecting uncertainties surrounding model parameters can drastically underestimate flood risks

Sanjib Sharma[1], Benjamin Seiyon Lee[2], Iman Hosseini-Shakib[1], Murali Haran[3], and Klaus Keller[4]

[1]Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA, USA
[2]Department of Statistics, George Mason University, Fairfax, VA, USA
[3]Department of Statistics, Pennsylvania State University, University Park, PA, USA
[4]Thayer School of Engineering at Dartmouth College, Hanover, NH, USA

Corresponding author: Sanjib Sharma (svs6308@psu.edu)

Key points:

- Current approaches to characterize flood hazards often sample only a relatively small subset of the known unknowns such as uncertainties about hydrologic model parameters.
- We implement a sequential Monte Carlo particle-based approach to improve the characterization of uncertainties surrounding hydrologic model parameters.
- Improving the characterization of model parametric uncertainty improves projections of flood hazards and risks.

Abstract

Floods drive dynamic and deeply uncertain risks for people and infrastructures. Uncertainty characterization is a crucial step in improving the predictive understanding of multi-sector dynamics and the design of risk-management strategies. Current approaches to estimate flood hazards often sample only a relatively small subset of the known unknowns, for example the uncertainties surrounding the model parameters. This approach neglects the impacts of key uncertainties on hazards and system dynamics. Here we mainstream a recently developed method for Bayesian data-model fusion to calibrate a computationally expensive distributed hydrologic model. We compare three different calibration approaches: (1) stepwise line search, (2) precalibration or screening, and (3) the new Fast Model Calibrations (FaMoS) approach. FaMoS deploys a particle-based approach that takes advantage of the massive parallelization afforded by modern high-performance computing systems. We quantify how neglecting known unknowns can drastically underestimate extreme flood events and risks. Accounting for parametric uncertainty improves model performance metrics over the best estimate

parameters. Improving the characterization of model parametric uncertainty improves hindcasts and projections of flood risks.

1. Motivation and Introduction

Floods pose major risks to people and property (Alfieri et al., 2017; Wing et al., 2018; Winsemius et al., 2015). These risks are dynamic and deeply uncertain (Merz et al., 2010; Read & Vogel, 2015; Ruckert et al., 2019; Zarekarizi et al., 2020). It is important to characterize the uncertainties surrounding flood hazards in order to understand the multi-sector dynamics and to inform the design of risk-management strategies (Boulange et al., 2021; Chester et al., 2020; Liu & Merwade, 2018; Salas et al., 2018b; Wasko et al., 2021; Wong & Keller, 2017).

Hydrologic models are commonly used to understand hydrological processes, predict the response of hydrological systems to changing stresses, and provide boundary conditions to estimate flood hazards and risks (Bates et al., 2021; Brunner et al., 2020; Judi et al., 2018; Koren et al., 2004; Rajib et al., 2020; Thorstensen et al., 2016). However, hydrologic projections are subject to deep uncertainties (Beven, 2014; Fisher & Koven, 2020; Hu et al., 2019; Mendoza et al., 2015). Deep uncertainty refers to a situation where the system model and the input parameters to the system model are not known or widely agreed on by the experts and/or decision makers (Lempert, 2002). Many studies are mostly silent on the deep uncertainty surrounding the model parameter (parametric uncertainty). Parametric uncertainty can arise from the epistemic uncertainties about model parameters (for example due to divergent expert priors) and different choices of calibration approaches. Hydrologic models need to resolve the complex response of multiple processes (e.g., land surface characteristics, soil properties and climate variability) with strong nonlinear interactions and often few observations. Characterizing parametric uncertainty can be critical to improve prediction credibility and inform decision-making, for example, in the context of water-resources planning and flood-risk management (Herman et al., 2013; Ruckert et al., 2019; Wong & Keller, 2017; Zarekarizi et al., 2020).

Previous studies provide valuable new insights on flood hazard and risk estimates using model simulations (Bates et al., 2021; Judi et al., 2018; Rajib et al., 2020; Sanders et al., 2020; Sharma et al., 2021; Wing et al., 2018). For example, Judi et al. (2018) demonstrates an integrated multimodel multiscale simulation approach to evaluate social, economic, and infrastructure resilience to future flooding. Rajib et al. (2020) develops a coupled land surface hydrologic and river hydraulic

modeling framework to provide regional flood hazard and risk estimates. Bates et al. (2021) presents estimates of current and future flood risk for all properties in the conterminous United States using a combined modeling approach considering river, coastal, or rainfall flooding. These studies typically obtain an optimal parameter set that produces the best possible agreement between simulated and observed streamflow hydrographs at target locations. These previous studies break important new ground, but are mostly silent on the impacts of parametric uncertainties on hazards and dynamics. Neglecting parametric uncertainties can underestimate the tails of flood hazard probability distribution (Bates et al., 2021; Mendoza et al., 2015; Rojas et al., 2020; Salas et al., 2018a), and can result in poor decisions and outcomes (Ruckert et al., 2019; Wong & Keller, 2017; Zarekarizi et al., 2020).

Studies that calibrate hydrologic models often manually adjust a subset of model parameters (Bitew & Gebremichael, 2011; Siddique & Mejia, 2017). These manual calibrations typically rely on visual inspection of streamflow hydrograph and a trial and error-based procedure; hence, this method can be rather labor-intensive and time-consuming (Lahmers et al., 2021; Siddique & Mejia, 2017). A more complex approach adopted in this area is automatic parameter optimization (Kamali et al., 2013; Van Liew et al., 2005). Automatic calibration relies on systematic search approaches to find the best parameter values based on predefined single- and/or multi-objective functions (Kamali et al., 2013). Some studies use surrogate methods such as Gaussian process-based emulators to help identify best-fit parameters (Gou et al., 2020; Pianosi et al., 2016; Razavi & Tolson, 2013). Gou et al. (2020) presents an automatic calibration framework that combines sensitivity analysis and surrogate-based optimization for calibrating catchment-specific hydrologic model parameters. Surrogate-based methods are typically limited to cases with relatively fewer model parameters because training a surrogate model can be computationally prohibitive with high-dimensional inputs due to the large number of training data required (Hwang & Martins, 2018; Lee et al., 2020; Liu & Guillas, 2017) or repeated evaluations of the gradient of the model output with respect to the input parameters (Constantine et al., 2014; Lataniotis et al., 2020).

Bayesian calibration of hydrologic models have become increasingly popular (Hsu et al., 2009; Jeremiah et al., 2011; Kavetski et al., 2018; Raje & Krishnan, 2012; Razavi & Tolson, 2013; Shafii et al., 2015; Su et al., 2018; Zhu et al., 2018). For example, Jeremiah et al. (2011) calibrates a conceptual water balance model by approximating the model parameters' posterior distribution

using adaptive Metropolis Markov chain Monte Carlo (MCMC) samplers and sequential Monte Carlo methods. Su et al. (2018) uses a Bayesian hierarchical model to calibrate the Priestly–Taylor Jet Propulsion Laboratory model using observed evapotranspiration measurements. Given the relatively short model run times, the hierarchical model can be fit using the Differential Evolution Markov Chain (Braak, 2006; Storn & Price, 1997), a population MCMC algorithm. Zhu et al. (2018) calibrates eight parameters of a conceptual water balance model using a Particle Evolution Metropolis Sequential Monte Carlo (PEM-SMC). The PEM-SMC algorithm evaluates the water balance model 2, 000 times sequentially, which may be computationally prohibitive for distributed hydrologic models with longer run times. These studies break important new ground, but focus on calibrating (1) average response of process over the watershed using a lumped hydrological model; (2) limited number of model parameters; (3) low-to-moderate flow threshold; and (4) relatively small basins. However, the computational requirement can be drastically larger for fully distributed hydrological modeling over the large basin and with a large number of sensitive parameters.

Here we expand on previous studies and demonstrate an implementation of a Bayesian model calibration framework by: (1) considering a computationally expensive distributed hydrologic model; (2) taking advantage of the massive parallelization afforded by modern high-performance computing systems; (3) focusing on a large number of extreme streamflow events; (4) characterizing model parametric uncertainty, and (5) assessing the impacts of uncertainty characterization on projected flood-hazards and -risks.

2. Bayesian Model Calibration

Bayesian computer model calibration (Bayarri et al., 2007a; Higdon et al., 2004; Kennedy & O'Hagan, 2001; Sacks et al., 1989) typically addresses two main objectives: (1) to infer the input parameters (in other words: what is the best parameter estimates); and (2) to quantify the uncertainty underlying the parameters (in other words: what is the joint probability density function of the parameters). These parameter estimates are impacted by factors such as model-observation discrepancy (Bayarri et al., 2007b; Brynjarsdóttir & O'Hagan, 2014; Kennedy & O'Hagan, 2001) and measurement errors. The Bayesian model calibration framework (see the discussion in Kennedy and O'Hagan, 2001) facilitates both parameter estimation and uncertainty quantification while also accounting for external sources of uncertainty (e.g., discrepancy and

measurement errors). For each model parameter, we specify prior distributions based on expert knowledge and then update the priors by comparing the model runs to the observed data. The update proceeds by placing more weight on the parameter sets whose corresponding model runs align better with the observations. The resulting posterior (updated) distribution naturally provides both point and interval estimates of the model parameters in light of the newly acquired data. Let θ be the vector of the model parameters, σ^2 the variance of the (assumed) independent and identically distributed observational error, and δ the discrepancy term. The posterior distribution $\pi(\theta, \sigma^2, \delta | Z)$ is defined as:

$$\pi(\theta, \sigma^2, \delta | Z) \propto p(Z|\theta, \sigma^2, \delta) \times p(\theta) \times p(\sigma^2) \times p(\delta),$$

where $\pi(\cdot)$and $p(\cdot)$ denotes the probability density function of the posterior and prior distributions, respectively.

For complex deterministic models, the posterior distribution may not be available in closed form (Higdon, 2003; Oakley, 2009). In this case, a common approach is to approximate the posterior via sampling approaches such as Markov chain Monte Carlo (MCMC) or Sequential Monte Carlo. The choice of sampling approaches in influenced by several factors including: (1) the computational time requirements for a single model evaluation; (2) the number of model parameters to be calibrated, (3) the degree to which the algorithm can be parallelized, (4) the available computation environment, and (5) the available time for the computations. Markov chain Monte Carlo methods with the true model can be an excellent choice for models with short single model run times (Asher et al., 2015; Gramacy, 2020; Lee et al., 2020). Emulation-calibration approaches replace the hydrologic model with a faster surrogate model, or emulator, and then sample from the posterior distribution via MCMC. However, it may be computationally expensive to construct a high-fidelity surrogate model with many input parameters due to the large amount of training data needed to fully explore the input space (Gramacy, 2020; Liu & Guillas, 2017). In hydrological applications, emulation-calibration methods are often used to calibrate broader summaries of water resources (e.g., long-term water balance and hydroclimatology of a region) as the observed data; as opposed to the fine-scale spatiotemporal hydrodynamic processes (Liu et al., 2018) inherent to flood-modeling applications. Sequential Monte Carlo methods (SMC) (Kalyanaraman et al., 2016; Kantas et al., 2014; Lee et al., 2020; Morzfeld et al., 2018; Papaioannou et al., 2016) methods are a practical alternative approach for calibrating high dimensional models with a larger number of input parameters.

2.1. The Fast Model Calibrations (FaMoS) approach

The **Fa**st **M**odel Calibration**s** (FaMoS) approach (Lee et al., 2020) approximates the posterior distribution of the model parameters using a series of sampling, reweighting, and re-sampling steps. The basic premise of sampling-importance resampling (Gordon et al., 1993) is to draw independent samples from the model parameters' prior distribution and retain the parameter sets whose corresponding outputs closely resemble the actual observations. We choose the appropriate parameter sets using weights, typically based on a goodness-of-fit metric such as the log likelihood function. The parameter sets whose model outputs fit the observed data well are given larger weights and those that do not are assigned smaller weights. The (importance) weights $w(\theta)$ are defined as:

$$w(\theta) = \frac{f(\theta)}{q(\theta)} = \frac{\pi(\theta|Z)}{p(\theta)}, \quad (1)$$

where $f(\theta)$ is the target function and $q(\theta)$is the importance function. In this context, we specify the target function as the posterior distribution of the model parameters $\pi(\theta|Z)$ and importance function as the prior distribution of the parameters $p(\theta)$. We approximate the posterior distribution using the weighted empirical distribution $\tilde{\pi}(\theta|Z)$ defined as:

$$\pi(\theta|Z) \approx \tilde{\pi}(\theta|Z) = \sum_{i=1}^{N} w(\theta_i)\delta(\theta_i), \quad (2)$$

where $w(\theta_i)$is the importance weight and $\delta(\theta_i)$is a Dirac measure at $\theta_i$for the i-th sample.

In the fast particle-based approach (Lee et al. 2020), we draw an initial ensemble of model parameters (particles) from the prior distribution (i.e., importance function) and approximate the posterior distribution (target function) using the initial ensemble. When there is very little overlap in the high-probability regions of the prior and posterior distribution, the initial ensemble may not adequately approximate the posterior distribution due to: (1) weight degeneracy, where the vast majority of particles have near-zero weights; and (2) sample impoverishment, where we "resample" the existing particles based on the weights, and we are left with multiple copies of a few unique particles.

The FaMoS (Lee et al, 2020) mitigates these issues by gradually building up to the posterior distribution, a technique from iterated batch importance sampling (Chopin, 2002) and Sequential Monte Carlo. Here, we consider a series of intermediate posterior distributions where those earlier in the series closely resemble the prior distribution and those at the latter part better resemble the full posterior distribution. In the first cycle, we use particles from the prior distribution to

approximate an earlier intermediate posterior distribution. In the subsequent cycles, we use samples from an intermediate posterior distribution to approximate a later intermediate posterior distribution. We end the algorithm when the target distribution is the final posterior distribution. For cycles t=1,...,T, the t-th intermediate posterior distribution is:

$$\pi_t(\theta|\, Z) \;\propto\; p(Z|\theta)^{\gamma_t} \;\times\; p(\theta), \qquad (3)$$

where $\gamma_t$denotes the incorporation factor such that $0 = \gamma_0 \leq \gamma_1 \leq \ldots \leq \gamma_{T-1} \leq \gamma_T = 1$. Note that the 0-th intermediate posterior distribution ($\pi_0(\theta|\, Z)$) is simply the prior distribution $p(\theta)$with incorporation factor $\gamma_0 = 0$. Likewise, the T-th intermediate posterior distribution ($\pi_T(\theta|\, Z)$) is the full posterior distribution since γ_T =1.

At the end of each cycle, there still may be many replicates of a few unique particles, or sample impoverishment. To increase the number of unique particles, we "jitter" or "mutate" the particles through a carefully constructed kernel function (Gilks & Berzuini, 2001; Li et al., 2014; Liu & West, 2001). Upon completion of the fast particle-based calibration algorithm, we are left with an ensemble of updated parameter settings (particles) which sensibly approximate the posterior distribution. Lee et. al. (2020) also provides guidelines for choosing the number of cycles, how to mutate the particles, and how to construct these intermediate posterior distributions. We approximate the posterior distribution using "mutated" samples from the final (T-th) intermediate posterior distribution:

$$\pi(\theta|Z) = \pi_T(\theta|Z) \;\approx \sum_{i=1}^{N} w_T\,(\widehat{\theta}_i)\delta(\widehat{\theta}_i) \qquad (4)$$

where $\hat{\theta}_i$ is the i-th mutated particle, $w_T(\hat{\theta}_i)$ are the corresponding weights from the T-th cycle, and $\delta(\widehat{\theta}_i)$is a Dirac measure at $\hat{\theta}_i$. We provide technical details about FaMoS in the Appendix.

3. Experimental Design

We demonstrate the approach for a case study in the Susquehanna River basin, Pennsylvania, United States. Pennsylvania provides a relevant study area as it ranked second, tenth, and fourteenth in the United States in terms of the frequency of flash flood-related fatalities, injuries, and casualties in 1959-2005 (Ashley & Ashley, 2008). This region has experienced several devastating flooding events over the recent decades, including floods associated with the remnants of Hurricane Ivan (September 2004), late winter–early spring extratropical systems (April 2005), warm-season convective systems (June 2006), and tropical storm Lee (September 2011) (Gitro et al., 2014; Grumm, 2011). In Pennsylvania, the Federal Emergency Management Agency (FEMA)

paid $953 million in property damages to National Flood Insurance Program participants between 1975 and 2019 (FEMA, 2019).

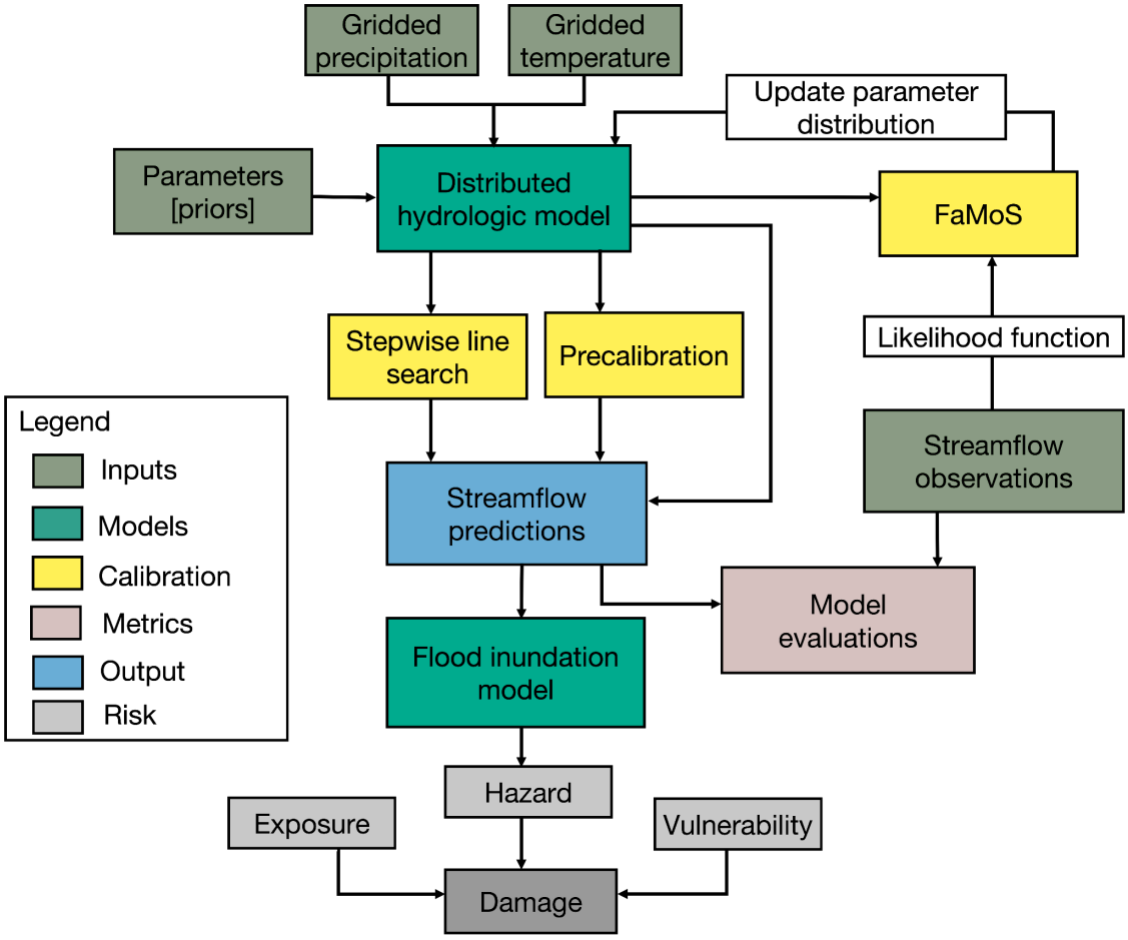

Figure 1: Diagrammatic representation of distributed hydrological model calibration framework. The framework also demonstrates flood hazards and risk components.

We use the National Oceanic and Atmospheric Administration's (NOAA) Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) (Koren et al., 2004). We run HL-RDHM in a fully distributed mode at a spatial resolution of 2 km. The 2 × 2 km^2 resolution mainly allows for a more realistic representation of the stream network. Within HL-RDHM, we use the Sacramento Soil Moisture Accounting model with Heat Transfer (SAC-HT) (Koren et al., 2004) to represent hillslope rainfall-runoff processes, and the SNOW-17 module (Anderson et al., 2006) to represent snow accumulation and melt. SAC-HT is a physics-based, conceptual model where the basin system is divided into regularly spaced, square grid cells to account for spatial

heterogeneity and variability. Each grid cell, in turn, is composed of storage components that store and transmit water. The cells are ultimately connected to each other through the stream network system, that is, each cell acts as a hillslope capable of generating surface and subsurface runoff that discharges directly into the streams. The hillslope runoff, generated at each grid cell by the SAC-HT and SNOW-17, is routed to the stream network using a nonlinear kinematic wave algorithm (Koren et al., 2004). Further information about the HL-RDHM model can be found for example in Koren et al. (2004), Reed et al. (2004), and Anderson et al. (2006).

We use three main datasets: multisensor precipitation estimates, gridded near-surface air temperature, and streamflow. We use NOAA's multisensor precipitation estimates and gridded near-surface air temperature products to run the hydrological model for parameter calibration purposes and to initialize the model. Multisensor precipitation estimates represent a continuous time series of hourly, gridded precipitation observations at $4 \times 4\ km^2$ cells, which are produced by combining multiple radar estimates and *in situ* rain-gauge measurements (Prat & Nelson, 2015; Rafieeinasab et al., 2015). The gridded near-surface air temperature data are derived by combining multiple temperature observation networks, including the meteorological terminal aviation routine weather report (METAR), USGS stations, and National Weather Service Cooperative Observer Program (Siddique & Mejia, 2017). We use streamflow observations from the United States Geological Survey gage 01554000 located at Susquehanna River at Sunbury, Pennsylvania. The selected gage station represents the drainage area of 47, 396 km^2.

We calibrate the model for the period of 2004-2008 and use 2009-2012 observations to evaluate the calibration performance. We use the year 2003 to spin up the model. As part of the calibration process, we select 12 out of the 17 model parameters associated with each model grid cell (Table S1). We only consider the model parameters that have a strong influence on the model output (see Figure S1). Exploring a higher-dimensional parameter space demands additional processors (particles) (Bain & Crisan, 2008; Jeremiah et al., 2011; Kantas et al., 2014) to sensibly calibrate the hydrological model. Selecting only the strongly influential model parameters can help reduce the computational costs considerably. This is, of course, an approximation and points to future research needs. The sensitive parameters are associated with different hydrodynamic processes related to baseflow, percolation, evaporation, snowfall, storm runoff, and channel routing (Table S1). These parameters are also suggested by several other studies (Gomez et al.,

2019; Sharma et al., 2021; Siddique & Mejia, 2017; Zarzar et al., 2018) as the most sensitive parameters in the Susquehanna river basin.

We compare Bayesian calibration with relatively simple and low-cost model calibration approaches: i) stepwise line search (Kuzmin et al., 2008) and ii) precalibration (Edwards et al., 2011). Stepwise line search typically adjusts a subset of model parameters to minimize an objective function (e.g., root mean square error) and returns a single estimate of the model parameters (for details of the implementation please see Text S2)(Bowman et al., 2017; Carlberg et al., 2020; Fares et al., 2014; Mejia & Reed, 2011; Siddique & Mejia, 2017). Precalibration applies a screening criterion to a large ensemble of hydrologic model runs and rules out any implausible model runs that deviate substantially from the observations (refer Text S3 for the details) (Craig et al., 1997; Edwards et al., 2011; Holden et al., 2010; Tarawneh et al., 2016).

We evaluate the calibrated model performance using several decision-relevant metrics. We use traditional deterministic metrics such as the Kling-Gupta Efficiency (KGE) (Mizukami et al., 2019), which provides a direct assessment of streamflow time series (e.g., shape, timing, water balance and variability) using the ensemble mean estimate. We also evaluate the probabilistic prediction skill using the Brier Skill Score (BSS) (Murphy, 1973) and the Continuous Ranked Probability Skill Score (CRPSS) (Murphy, 1970). The Brier score is essentially the mean squared error of the probability predictions, considering that the observation is one if the event occurs, and that the observation is zero if the event does not occur. The Continuous Ranked Probability Score measures the integral square difference between the cumulative distribution functions of the observation and predictions, averaged over all pairs of predictions and observations. The selection of these decision-relevant metrics is motivated by the balance between model output goodness-of-fit, calibration approaches, and data availability. The description of evaluation metrics is provided in Text S4 in the supporting information. The evaluation is focused on high flows by choosing the river flow that exceeds NOAA's Action Stage (McEnery et al., 2005). Action Stage refers to the stage which, when reached by a rising river, represents the level where the National Weather Service or a partner/user needs to take some mitigation action in preparation for possible significant hydrologic activity.

We assess the impact of model calibration on flood damage estimates. Flood damage represents interactions among hazard, exposure and vulnerability (Tellman et al., 2021; Wing et al., 2018). Hazard in this case refers to the magnitude of the flood event. Exposure characterizes

property value in the floodplain. Vulnerability characterizes how sensitive the impacts are for a given hazard and exposure. We consider 2,000 hypothetical houses to quantify the damage from flood hazards (Figure S4; TextS6). We assess damage for a certain depth of water in a house by using a relatively simple Bathtub-based flood inundation model (Didier et al., 2019; Fereshtehpour & Karamouz, 2018; Neumann & Ahrendt, 2013; Yunus et al., 2016) and a vulnerability model (Scawthorn et al., 2006). The Bathtub model relies on a digital elevation model to provide flood depth in a house for a particular corresponding water level in the river (refer TextS5 and TextS6 for the details). We use a common vulnerability model (depth-damage function) provided by the Federal Emergency Management Agency (FEMA) (Scawthorn et al., 2006).

4. Results and Discussion

We first generate streamflow simulations using the "best" parameter estimates obtained via the stepwise line search (Figure 2). In the considered example, stepwise line search substantially underestimates the high streamflow (Figure 2). Stepwise line is designed to sample high-probability outcomes and excludes comprehensive sampling of the parametric distribution (Kuzmin et al., 2008; Sharma et al., 2019).

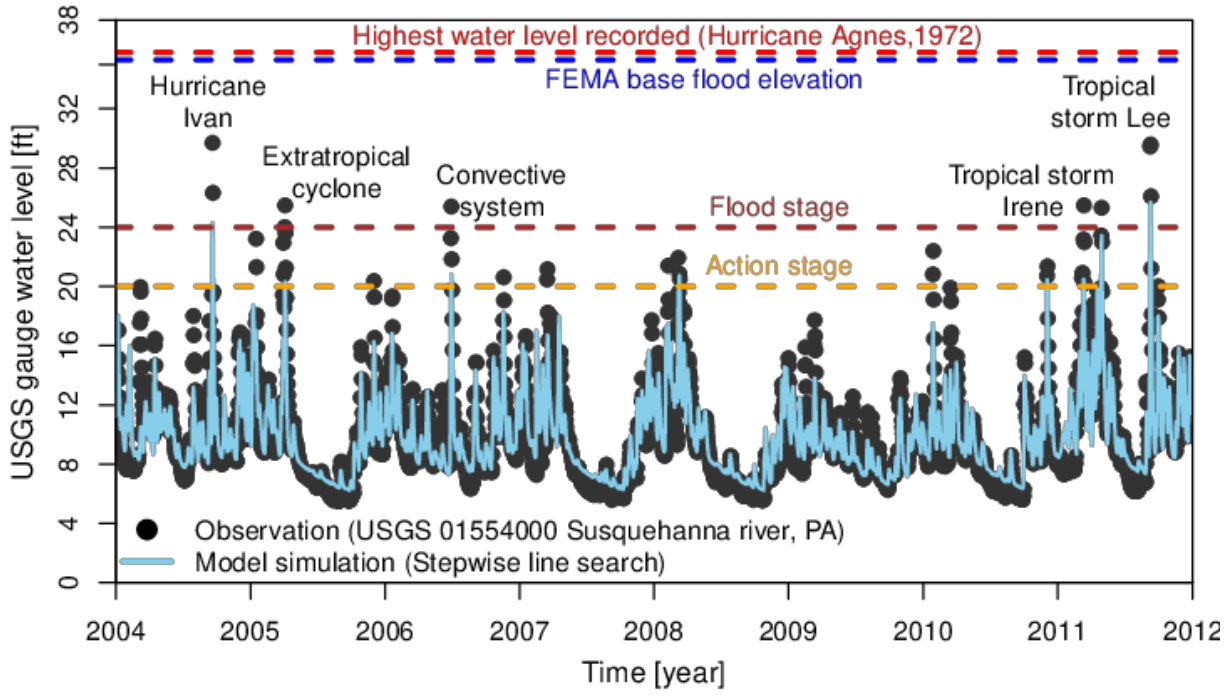

Figure 2: Historical time series of water level observation and model simulations obtained using best parameter estimates (stepwise line search). We obtain the observation from the United States Geological Survey (USGS)

gauge records for ID 01554000 located upstream of Selinsgrove, Pennsylvania, USA. The most destructive floods in the Susquehanna river basin that occurred in recent years, each associated with different flood-generating mechanisms, includes Hurricane Ivan (September 2004), late winter–early spring extratropical systems (April 2005), warm-season convective systems (June 2006), and tropical storm Lee (September 2011).

We account for parametric uncertainty using precalibration and FaMoS (Figure S1). Characterizing parametric uncertainty requires knowledge of model behavior throughout the (often high-dimensional) parameter space. Precalibration provides a relatively simple method to explore the high-dimensional parameter space. Precalibration is a low-cost way of ruling out implausible model runs. We begin with an initial ensemble of 5,000 model runs with input parameters settings selected from a 12-dimensional Latin hypercube design (Helton & Davis, 2003). We select an ensemble of 165 runs that fall within the +/- 75% window surrounding each observation. Note that specifying bounds for precalibration is a subjective choice (Craig et al., 1997; Edwards et al., 2011; Holden et al., 2010; Tarawneh et al., 2016). This choice impacts the "surviving" parameter samples. For instance, imposing tight bounds on the observed streamflow could lead to high-resolution sampling of the plausible parameter space and wider bounds may include more implausible runs into the final ensemble. We choose the considered acceptable range to sample into the upper tails of projected flood hazards, which are often associated with high-cost events.

FaMoS adopts a more complex (but also more powerful) calibration approach compared to precalibration. We incorporate domain-area expertise (prior distribution) of the unknown parameters and also account for additional sources of uncertainty such as model-observation discrepancies and observational error (see the Appendix for the details). As a result, we obtain a distribution of viable parameter values (posterior distribution) along with interval estimates, as opposed to a single best fit estimate (Figure S1). Unlike precalibration, FaMoS does not fix an arbitrary screening criterion, but rather uses a flexible statistical model to assess model-fit. Moreover, FaMoS sequentially explores the entire parameter space and systematically attempts to move to a "target" region that contains the most plausible sets of model parameters. In contrast, precalibration attempts to locate this "target" region using a single initial ensemble of model runs.

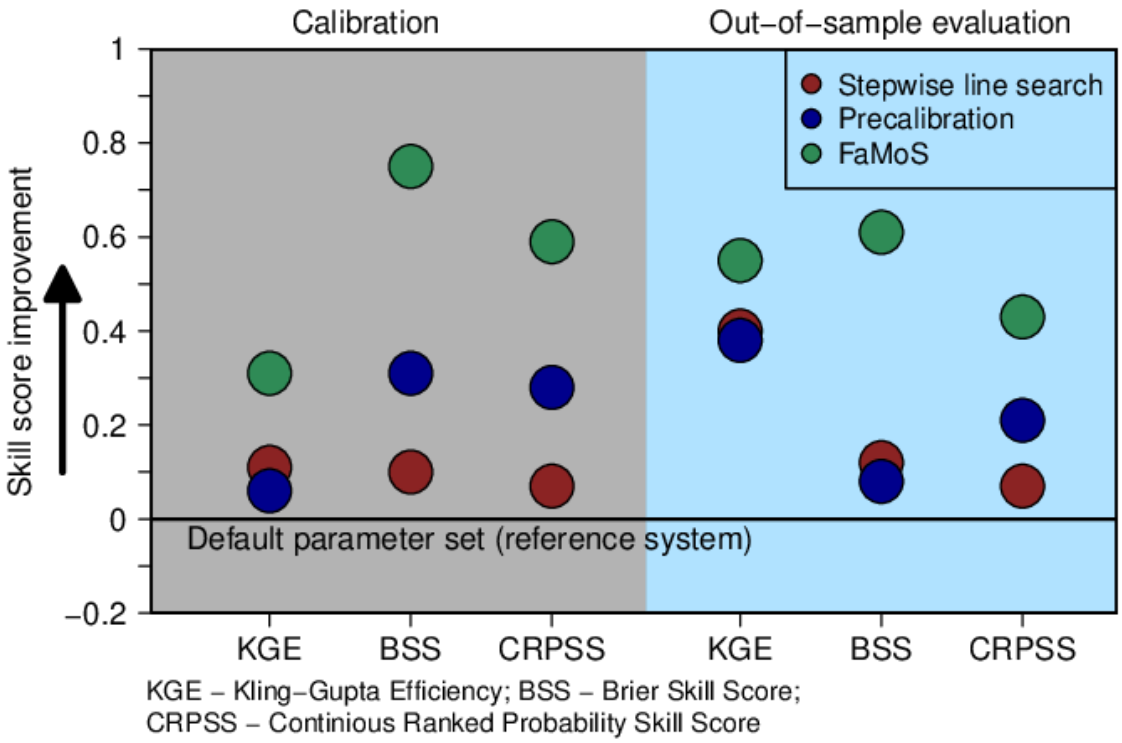

Figure 3: Performance metrics for hydrological model calibration and out-of-sample prediction. We compute Kling-Gupta Efficiency (KGE), and Brier skill score (BSS), and Continuous ranked probability skill score (CRPSS). All the metrics are computed with reference to the default parameter set available from several previous studies (Anderson et al. 2006, Reed et al. 2004). Any positive values of the skill score, from 0 to 1, indicate that the calibration approach performs better than the reference system. Thus, a skill score of zero indicates no skill, and a skill of one indicates perfect skill.

Accounting for parametric uncertainty improves model performance metrics for the calibration data and out-of-sample predictions (Figure 3). We compute the skill score (KGE, BSS, and CRPSS) with reference to raw (uncalibrated) model runs using default parameter estimates obtained from several previous studies (Anderson et al., 2006; Reed et al., 2007). In terms of the performance metrics, model predictions remain skillful for all the calibration approaches (Figure 3). Precalibration outperforms the stepwise line search (best estimate predictions). FaMoS demonstrate a higher skill score than both the stepwise line search and precalibration for both calibration and out-of-sample evaluations.

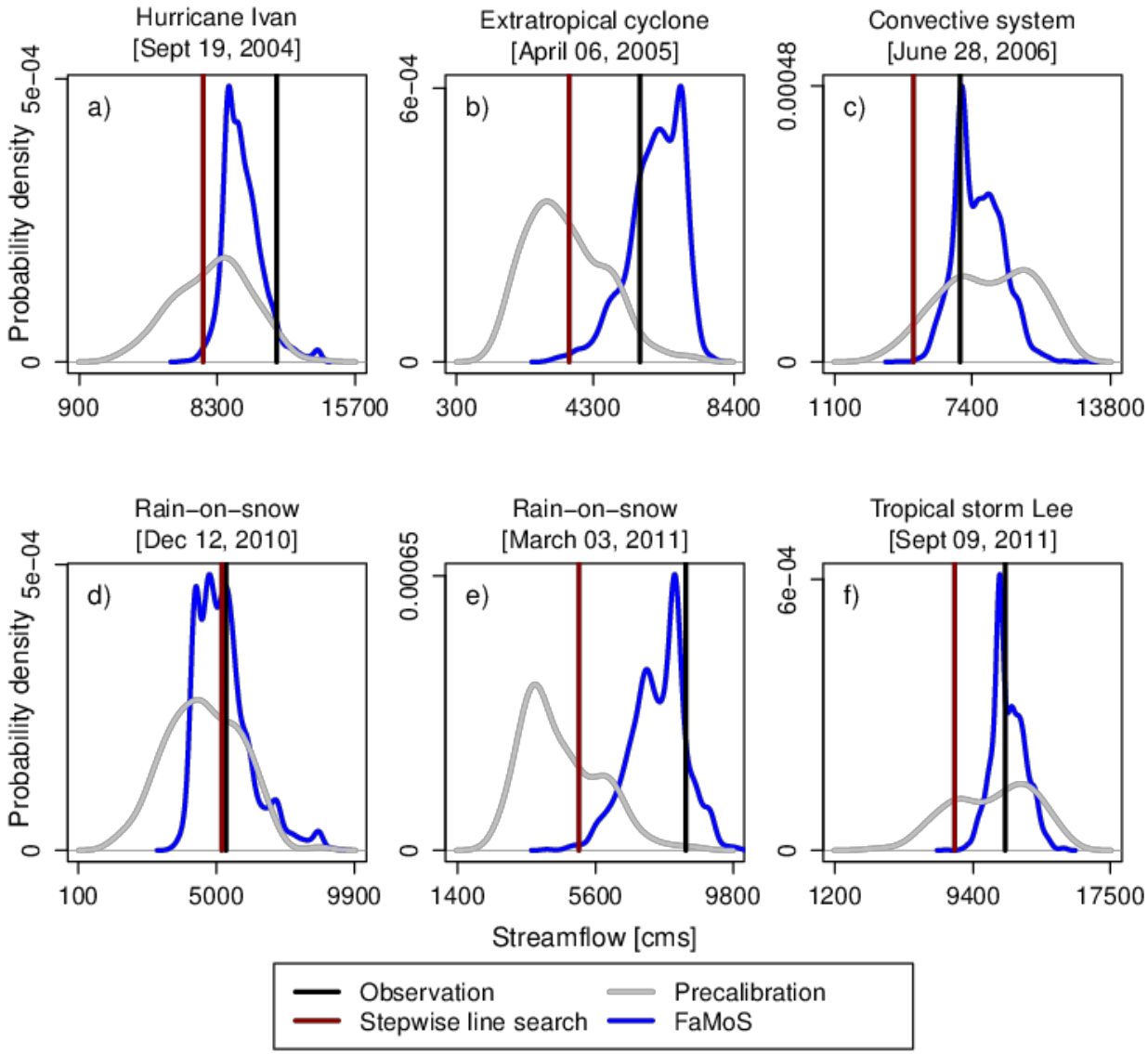

Figure 4: (a) - (c) Calibration and (d) - (f) and out-of-sample prediction for different flood events.

Accounting for parametric uncertainty improves flood hazard estimates (Figure 4). The resulting predictive distribution of flood events demonstrates the impacts of model calibration. The stepwise line search underestimates the flood peaks by as much as 35% (Figure 4b) during calibration and 40% during out-of-sample prediction (Figure 4e). Precalibration captures the specific flood events, but exhibits very high prediction uncertainty as evidenced by the wider prediction intervals. Overall, FaMoS improves flood peak estimates and provides narrower prediction intervals. Consider, as an example, the case of Tropical Storm Lee with streamflow

observation of 11, 292 m^3/sec. Precalibration provides a flood peak prediction of 10, 539 m^3/sec and prediction intervals (5%-95% credible interval) range from 6, 359 m^3/sec to 14, 222 m^3/sec (width = 7, 863 m^3/sec). FaMoS has a corresponding flood peak prediction of 11, 467 m^3/sec with a credible interval ranging from 9, 925 m^3/sec to 13, 121 m^3/sec (width = 3, 196 m^3/sec).

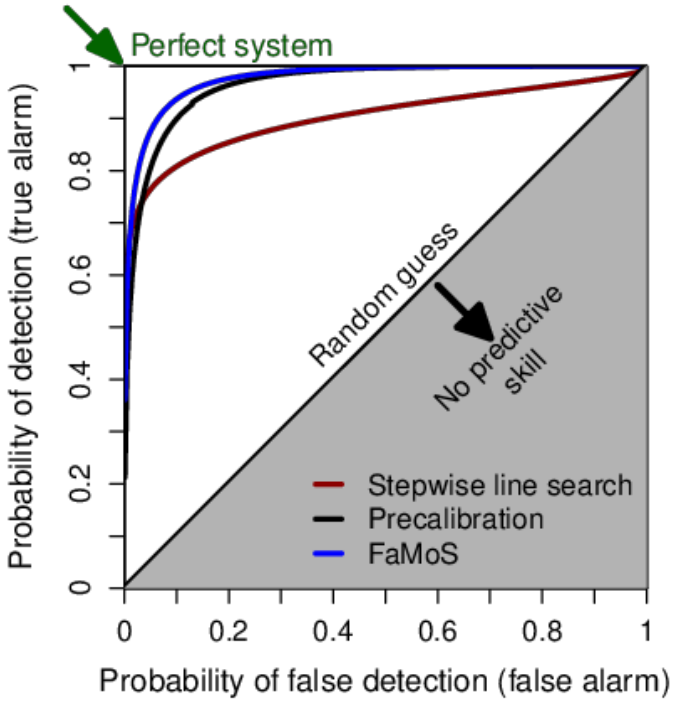

Figure 5: Relative operating characteristics (ROC) curve for different calibration approaches. ROC curve plots the probability of detection against the probability of false detection for a range of forecast probability levels. A larger area under the ROC curve represents a more skillful prediction, with more ability to discriminate between flood thresholds. The area under the ROC curve can range between 0 and 1, where a score of 1 implies perfect discrimination and a score of 0.5 or less implies predictive discrimination that is no better than a random guess. We also compute the ROC score. The ROC score measures the average gain over climatology for all probability levels. The ROC score for stepwise line search, precalibration and FAMOS is 0.55, 0.85 and 0.96 respectively.

We assess each calibration approach's classification ability or how well each method discriminates between occurrences (water level crossing the action stage) versus non-occurrences (regular water level) of an event (Figure 5). Managing flood risks can require decision makers to choose between two options (e.g., to evacuate or not or to elevate a house or not) based on a prediction of an event (e.g., water rising to a certain level) with one decision preferred if the event

doesn't occur, and the other if it does. A perfect prediction system for a binary outcome correctly predicts the occurrence of an event (unity probability of detection) and never issues incorrect predictions when it does not occur (zero probability of false detection). How well a prediction system approaches this ideal case can be quantified by the relative operating characteristics (ROC) curve (see Text S4) (Mason & Graham, 2002). Technically, the ROC curve assesses the quality of probability predictions by relating the probability of detection (true alarm) to the corresponding probability of false detection (false-alarm rate), as a decision threshold is varied across the full range of a continuous prediction quantity (Figure 5). Streamflow predictions obtained using the FaMoS parameter distribution exhibit better discriminatory ability (higher ROC score) than the stepwise line search and precalibration. Stepwise line search shows a relatively poor ability to discriminate between different events. This poor ability to discriminate between the events can lead to poor decisions and outcomes.

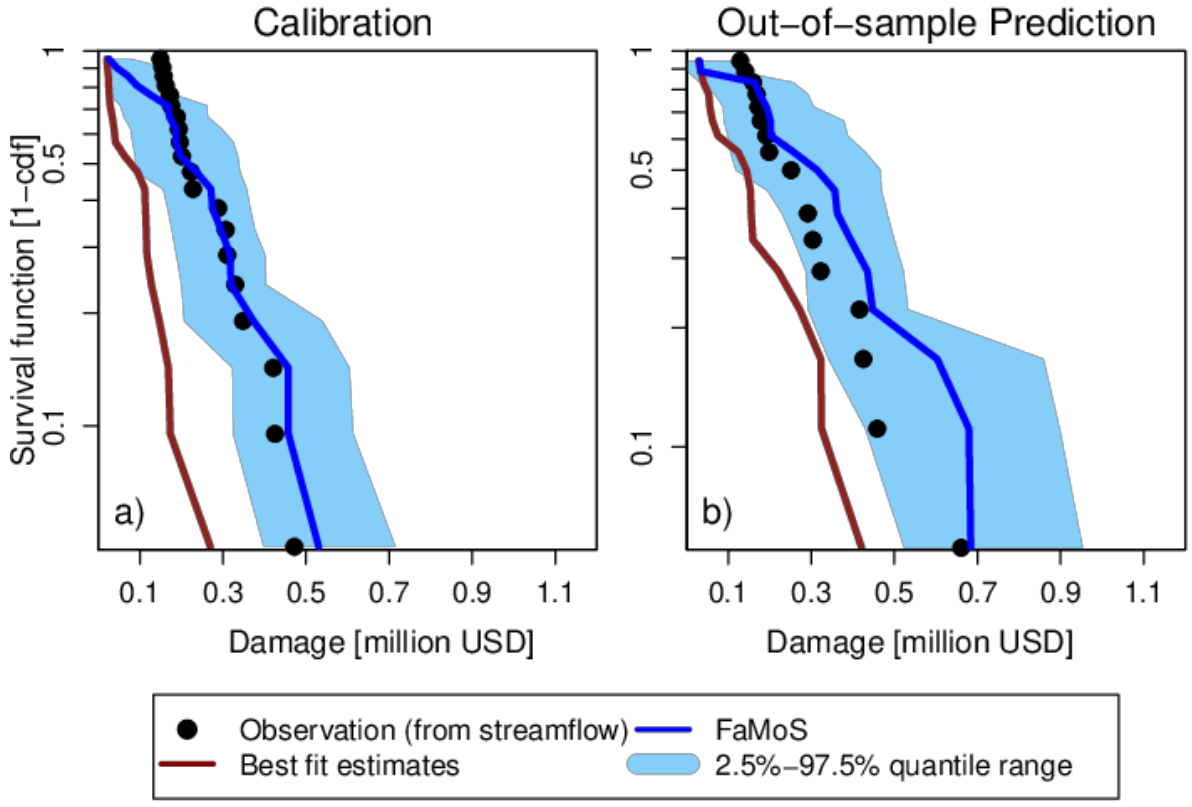

Figure 6: Survival function (one minus the cumulative frequency) for damage estimates using streamflow obtained using the best parameter set (stepwise line search) and parameter distribution (FaMoS). We shoe damage estimates for a) calibration and b) out-of-sample prediction. cdf= cumulative distribution function.

Neglecting parametric uncertainty also underestimates potential flood damage (Figure 6). We find that the stepwise line search tends to underestimate the flood damage. The underestimation bias increases as flood magnitude increases. Accounting for parametric uncertainty improves the damage estimates for the calibration data and out-of-sample predictions. The damage credible interval obtained using FaMoS parameter distribution generally captures the observed damage for different flood events. As expected, at the upper tails of the damage, the predictive uncertainty tends to be higher for the out-of-sample prediction as compared to the calibration.

5. Caveats

We use a relatively simple model and small region with hypothetical exposure to demonstrate our points. This parsimony helps with transparency, but it comes with several caveats. For example, our analysis focuses on high flows. Future work might consider calibrating other flow thresholds, including low flows and moderate flows. Due to a large number of low and moderate flow observations, dimension-reduction techniques like principal components (Chang et al., 2014; Higdon et al., 2008) or eigenfunctions (Mak et al., 2018) may be appropriate to summarize the large datasets. There are, of course, other deep uncertainties affecting flood hazards and risks that could be taken into account in future work. These include model structural uncertainty as well as different spatial resolutions and land surface characteristics. Increasing the spatio-temporal resolutions may drastically raise the hydrologic model's complexity as well as the associated single model run times. To reduce the number of sequential hydrologic model evaluations, we can embed parallel Markov Chain Monte Carlo approaches such as Multiple-Try Metropolis (Liu et al., 2000) or "emcee" samplers (Goodman & Weare, 2010)or genetic algorithms (Park et al., 2009) into the FaMoS calibration framework. We note that our damage estimates are based on a simple Bathtub-based flood inundation model. Future work could use process-informed models to characterize the impacts of hydrodynamic processes in damage estimates (Brunner, 1995; Coulthard et al., 2013; Judi et al., 2018). In addition, future work could sample the uncertainty surrounding the flood vulnerability of the building (Wing et al., 2020).

6. Conclusions

We use a Bayesian data-model fusion framework to calibrate a distributed hydrologic model and to demonstrate practical implications of neglecting key uncertainties on hazard- and

risk-estimates. We compare the results of the Bayesian approach to two simpler methods: stepwise line search and precalibration. We show that these simpler methods can considerably underestimate flood hazards and risks. Precalibration improves flood hazards estimates over the best fit estimates, but provides a wider predictive interval (i.e., highly uncertain estimates). The predictive skill of the Bayesian approach dominates the stepwise line search and precalibration approaches. We show how neglecting model parametric uncertainty can substantially underestimate flood hazards and risk estimates and demonstrate how applying state-of-the-art statistical methods can help to refine flood-risk projections.

Acknowledgments

This study was co-supported by the US Department of Energy, Office of Science through the Program on Coupled Human and Earth Systems (PCHES) under DOE Cooperative Agreement No. DE-SC0016162 and DE-SC0022141 as well as the Penn State Center for Climate Risk Management. We thank Rob Nicholas, Skip Wishbone, Dave Judi, and the PSIRC team for inputs. All errors and opinions (unless cited) are those of the authors and not of the funding entities.

Disclaimer and License

The results, data, software tools, and other resources related to this work will be available under the GNU general public open-source license, as-is, without warranty of any kind, expressed or implied. In no event shall the authors or copyright holders be liable for any claim, damages, or other liability in connection with the use of these resources. This is academic research and not designed to be used to guide a specific decision.

Author contributions

All authors contributed to the study design. S.S. led the hydrologic analysis. B.L. and M.H. constructed the particle-based calibration model. I.H.S led the flood damage analysis. I.H.S. performed a code review. S.S., B.L, and K.K wrote the initial draft of the manuscript. All authors revised and edited the manuscript.

Data and Code Availability

The code used for this analysis and the data required to plot the results is available through a publicly accessible GitHub repository and under the GNU open-access license upon acceptance to a peer-reviewed journal. Reviewers can access these resources from https://github.com/benee55/FamosHydroModel. All data and code currently available at GitHub will be published via Zenodo upon article acceptance.

Competing interests

The authors are not aware of any competing financial or nonfinancial interests.

Materials & Correspondence

Correspondence and requests for materials should be addressed to the corresponding author. The code and data are available on GitHub (made public upon acceptance of the paper).

Appendix A: Fast Model Calibrations (FaMoS) Details

1 Bayesian Calibration Framework

Suppose we have an observed time series $\mathbf{Z} = (Z(t_1), ..., Z(t_n))'$ times $t_i \in \mathcal{T}$ where $\mathcal{T}$ is the temporal domain of the process. We also have a deterministic computer model that generates a temporal process, or time series, at times $t_i \in \mathcal{T}$. Let $Y(t, \boldsymbol{\theta})$ be the computer model output at the time $t \in \mathcal{T}$ and the parameter (input) setting $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Θ is the parameter space of the computer model with integer d being the number of input parameters. In this study, we use a discretized temporal domain at T distinct time points $\mathcal{T} = (t_1, ..., t_T)'$. The vector $\mathbf{Y}(\boldsymbol{\theta}_i) = (Y(t_1, \boldsymbol{\theta}_i), ..., Y(t_T, \boldsymbol{\theta}_i))'$ is the computer model output corresponding to parameter setting $\boldsymbol{\theta}_i$. For input parameter setting $\boldsymbol{\theta}$, we model the observations $\mathbf{Z}$ as:

$$\mathbf{Z} = \mathbf{Y}(\boldsymbol{\theta}) + \boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{A1}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$ are the independently and identically distributed observational error, and $\boldsymbol{\delta} \in \mathbb{R}^n$ is a systemic data-model discrepancy term, which can be modeled as a zero-mean Gaussian process (Bhat et al., 2010; Bayarri et al., 2007) or other flexible functional forms (Brynjarsdottir and O'Hagan, 2014).

In the Bayesian calibration framework, we obtain samples (via a Markov chain Monte Carlo (MCMC) algorithm) from the posterior distribution:

$$\pi(\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta}|\mathbf{Z}) \propto p(\mathbf{Z}|\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta}) p(\boldsymbol{\theta}) p(\sigma_\epsilon^2) p(\boldsymbol{\delta}), \tag{A2}$$

where $p(\mathbf{Z}|\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta})$ denotes the likelihood function and $p(\cdot)$ represents the prior distribution for the respective parameters and discrepancy term. Note that each evaluation of $p(\mathbf{Z}|\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta})$ requires running the computer model using specific input parameters $\boldsymbol{\theta}$. Hence, MCMC-based calibration approaches are sensible for computer models with shorter single model run walltimes, typically under 5 seconds per model run (Lee et al., 2020). For our study, we estimate that a standard MCMC-based calibration approach would on the order of years to approximate the posterior distribution $\pi(\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta}|\mathbf{Z})$.

Surrogate methods such as Gaussian process-based emulators are well suited to computer models with long run times and few model parameters. Here, we replace the more expensive computer model with a cheaper emulator. Since an emulator must be trained using a pre-specified set of inputs and model output, a carefully designed set of traning data is important for building accurate surrogate models. Dense sampling schemes, such as full factorial or fractional factorial designs, capture higher order interactions; however, running the computer model at each of the design points is costly. Space-filling designs such as the Latin Hypercube Design (McKay et al., 2000; Steinberg and Lin, 2006; Stein, 1987) or adaptive experimental designs (Chang et al., 2016; Gramacy and Apley, 2015; Urban and Fricker, 2010; Queipo et al., 2005) use fewer design points, but these may generate lower-fidelity surrogate models by ignoring higher order interactions among inputs (Liu and Guillas, 2017).

2 Particle-based Calibration Framework

We calibrate the HL-RDHM distributed hydrological model using the fast particle-based approach from Lee et al. (2020), which is built upon traditional Sequential Monte Carlo algorithms (Del Moral et al., 2006; Doucet et al., 2000; Liu and West, 2001), notably the Iterated Batch Importance Sampling (IBIS) (Chopin, 2002; Crisan and Doucet, 2000) method. This method approximates a the posterior distribution $\pi(\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta}|\mathbf{Z})$ using an evolving ensemble of particles.

We simplify the notation for an arbitrary target distribution as $\pi(\boldsymbol{\theta})$ with random variable $\boldsymbol{\theta} \in \mathbb{R}^d$. In the calibration framework, the target distribution $\pi(\boldsymbol{\theta})$ would be the posterior distribution $\pi(\boldsymbol{\theta}, \sigma_\epsilon^2, \boldsymbol{\delta}|\mathbf{Z})$ with random variables $\boldsymbol{\theta}, \sigma_\epsilon^2$, and $\boldsymbol{\delta}$ and observations $\mathbf{Z}$. Suppose we want to estimate $\mu = E_\pi\big[g(\boldsymbol{\theta})\big]$. Given $q(\boldsymbol{\theta}) > 0$ whenever $g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) > 0, \quad \forall \boldsymbol{\theta} \in \Theta$. Then $E_\pi\big[g(\boldsymbol{\theta})\big] = E_q\Big[g(\boldsymbol{\theta})w(\boldsymbol{\theta})\Big]$, where $w(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$ is the importance weight and $\sum_{i=1}^N w(\boldsymbol{\theta}_i) = 1$. The importance sampling estimator is $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^N g(\boldsymbol{\theta}_i)w(\boldsymbol{\theta}_i)$ and $\hat{\mu}_n \to \mu$ with probability 1 by the strong law of large numbers. For target distributions with an unknown normalizing constant, the weights can be normalized as follows:

$$\tilde{w}(\boldsymbol{\theta}_i) = \frac{w(\boldsymbol{\theta}_i)}{\sum_{j=1}^n w(\boldsymbol{\theta}_j)} = \frac{\pi(\boldsymbol{\theta}_i)/q(\boldsymbol{\theta}_i)}{\sum_{j=1}^n w(\boldsymbol{\theta}_j)} \tag{A3}$$

where $\sum_{i=1}^N \tilde{w}(\boldsymbol{\theta}_i) = 1$.

Sampling-Importance-Resampling (Gordon et al., 1993; Doucet et al., 2001) approximates a target distribution $\pi(\boldsymbol{\theta})$ with an empirical distribution of the particles $\hat{\pi}(\boldsymbol{\theta})$ from an importance function $q(\boldsymbol{\theta})$. The empirical distribution is defined as:

$$\pi(\boldsymbol{\theta}) \approx \hat{\pi}(\boldsymbol{\theta}) = \sum_{i=1}^N \tilde{w}(\boldsymbol{\theta}_i)\delta(\boldsymbol{\theta}_i), \tag{A4}$$

where $\tilde{w}(\boldsymbol{\theta}_i)$ are the normalized importance weights, $\delta(\boldsymbol{\theta}_i)$ is a Dirac measure that places unit mass at $\boldsymbol{\theta}_i$ and $\sum_{i=1}^N \tilde{w}(\boldsymbol{\theta}_i) = 1$.

Poor choices of importance functions can lead to inaccurate approximations of the target distribution (Doucet et al., 2000) where the bulk of the particles $\boldsymbol{\theta}_i$'s do not reside in the high-probability regions of the target distribution $\pi(\boldsymbol{\theta})$. Weight degeneracy occurs when the vast majority of the particles have near-zero importance weights. Multinomial resampling methods can combat weight degeneracy by eliminating the particles with very small important weights and replicating those with higher weights (Gordon et al., 1993; Doucet et al., 2000). After resampling, we reset all importance weights such that $w(\boldsymbol{\theta}_i) = 1/N$ and use the unweighted empirical distribution $\ddot{\pi}(\boldsymbol{\theta})$:

$$\ddot{\pi}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N N_i\delta(\boldsymbol{\theta}_i), \tag{A5}$$

where N_i is the number of replicates corresponding to particle $\boldsymbol{\theta}_i$ and $\sum_{i=1}^N N_i = N$. Extreme weight degeneracy, where very few particles have any significant weight, can lead to sample impoverishment where a few unique particles $\boldsymbol{\theta}_i$'s are heavily replicated in the re-sampling

step; hence, the empirical distribution $\ddot{\pi}(\boldsymbol{\theta})$ may poorly approximate the target distribution $\pi(\boldsymbol{\theta})$.

An alternative method mutates the replicated particles with samples from $K(\boldsymbol{\theta}_i^{(t-1)})$, the Metropolis-Hastings transition kernel (Gilks and Berzuini, 2001), whose stationary distribution is also the target distribution $\pi(\boldsymbol{\theta})$. Here we run J Metropolis-Hastings updates for each particle $\boldsymbol{\theta}_i$, for $i = 1, ..., N$. Alternative mutation schemes use genetic algorithms (Zhu et al., 2018) or different families of transition kernels, $K(\cdot)$ (Papaioannou et al., 2016; Murray et al., 2016). We set the jth sample drawn via MCMC as the mutated particle $\tilde{\boldsymbol{\theta}}_i$. Since $\tilde{\boldsymbol{\theta}}_i \sim \pi(\boldsymbol{\theta})$, the resulting empirical distribution $\tilde{\pi}(\boldsymbol{\theta})$ approximates the target distribution $\pi(\boldsymbol{\theta})$:

$$\pi(\boldsymbol{\theta}) \approx \tilde{\pi}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \tilde{\boldsymbol{\theta}}_i \delta(\tilde{\boldsymbol{\theta}}_i). \tag{A6}$$

Unfortunately, poor importance functions can result in severe sample impoverishment, which may require very long (and costly) mutation stages to provide an accurate representation of the target distribution (Li et al., 2014). Mixture approximations (Gordon et al., 1993) or kernel smoothing methods (Liu and West, 2001) can mutate or rejuvenate the replicated particles. However, these methods may not scale well to high-dimensional target distributions (Doucet et al., 2000).

2.1 Fast Particle-based Approach For Computer Model Calibration

In this study, we aim to approximate the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$ from a computationally efficient approach. The fast particle-based approach (Lee et al., 2020) utilizes a set of tempered, or intermediate, posterior distributions $\pi_t(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$ for $t = 1, ..., T$, which will act as both the importance functions and target distributions. Intermediate posterior distributions can be generated using likelihood tempering (Chopin, 2002; Neal, 2001; Liang and Wong, 2001) where the tth intermediate posterior distribution is defined as:

$$\pi_t(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z}) \propto p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2)^{\gamma_t} p(\boldsymbol{\theta}) p(\boldsymbol{\delta}) p(\sigma_\epsilon^2), \tag{A7}$$

where γ_t's are determined according to a schedule where $\gamma_0 = 0 < \gamma_1 < \cdots < \gamma_T = 1$. For each $\pi_t(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$, the likelihood component is a fractional power of the original likelihood $p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2)$. Using an adaptive incorporation schedule (Lee et al., 2020), we can select the appropriate $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, ..., \gamma_T\}$ within the calibration algorithm.

For cycle $t = 1$, we set the importance distribution to be the prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2) = p(\boldsymbol{\theta}) p(\boldsymbol{\delta}) p(\sigma_\epsilon^2)$, and the target distribution to be the first intermediate posterior distribution, $\pi_1(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$. For subsequent cycles t, the importance distribution is $\pi_{t-1}(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$ and the target distribution is $\pi_t(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$.

Next, we mutate the particles via short runs of the Metropolis-Hastings algorithm, where the stationary distribution is $\pi_t(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2|\mathbf{Z})$, the t-th intermediate posterior distribution. Note that the importance and target distributions are consecutive (t-th and (t+1)-th) intermediate posterior distributions, so there is considerable overlap between the high-probability regions of the two distributions. In the mutation stage, we employ the stopping rule from Lee et al.

(2020) to control the number of Metropolis-Hastings updates; thereby preventing any unnecessary computer model runs. The mutation stages ends when the Bhattacharyya distance (Bhattacharyya, 1946) between two sets of particles from the mutation stage stablizes.

2.2 Adaptive incorporation schedule

To reduce computational costs and potentially reduce unnecessary computer model evaluations, we adopt the adaptive incorporation schedule from Lee et al. (2020). Upon initialization, we set the first incorporation increment $\gamma_0 = 0$. We draw the initial set of particles $\boldsymbol{\theta}_0$ from $\pi_0(\boldsymbol{\theta}|\mathbf{Z}) \propto L(\boldsymbol{\theta}|\mathbf{Z})^0 p(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, the prior distribution of model parameters. For the subsequent cycles $t = 1, 2, 3, ...,$ we calculate the full likelihood $L(\boldsymbol{\theta}_{t-1}^{(i)}|\mathbf{Z})$ for $i = 1, ..., N$ where $\boldsymbol{\theta}_{t-1}^{(i)}$ denotes the parameter samples from the previous cycle $t-1$. For computational efficiency, we reuse the likelihood evaluations from the previous cycle. Next, we compute the optimal γ_t that returns an effective sample size (ESS) of ESS_{thresh} or a sample size closest to ESS_{thresh}: $\gamma_t = \text{argmin}_\gamma\{(ESS_\gamma - ESS_{thresh})^2\}$, where $\gamma \in (\gamma_{min}, 1 - \gamma_{t-1})$, γ_{min} is a previously set minimum incorporation value, $ESS_{\gamma_t} = \sum_{i=1}^{N} 1/w_t(\boldsymbol{\theta}_t^{(i)})^2$, and $w_t(\boldsymbol{\theta}_t^{(i)}) \propto L(\boldsymbol{\theta}_t^{(i)}|\mathbf{Z})^\gamma$. Note that we can lower computational costs by evaluating the full likelihood $L(\boldsymbol{\theta}_0^{(i)}|\mathbf{Z})$ only once before the optimization.

We stop the scheduling algorithm when $\sum_{i=1}^{t} \gamma_t = 1$, or when the entire likelihood has been incorporated and the target distribution evolves to the full posterior distribution $\pi(\boldsymbol{\theta}, \sigma_\epsilon^2|\mathbf{Z})$. Note at each cycle t, we set the incorporation increment (γ_t) to be between γ_{min} and $1 - \sum_{i=1}^{t} \gamma_t$. The user will typically set the minimum incorporation increment γ_{min} and the threshold effective sample size, ESS_{thresh}. We provide our choice of γ_{min} and ESS_{thresh} in the next section (Implementation Details).

Adaptive likelihood incorporation schedule

1. Initialization: At $t = 0$, set $\gamma_0 = 0$.
2. When $t > 0$ and $\sum_{i=1}^{t-1} \gamma_i < 1$
   - Compute $L(\boldsymbol{\theta}_{t-1}^{(i)}|\mathbf{Z})$ for $i = 1, ..., N$
   - Set $\gamma_t = \text{argmin}_\gamma\{(ESS_\gamma - ESS_{thresh})^2\}$, where $ESS_\gamma = \sum_{i=1}^{N} \frac{1}{w_t^{(i)2}}$, $w_t^{(i)} \propto L(\boldsymbol{\theta}_t^{(i)}|\mathbf{Z})^\gamma$, and $\gamma \in (\gamma_{min}, 1 - \gamma_{t-1})$.
   - γ_{min} is a predetermined minimum incorporation value
3. When $\sum_{i=1}^{t-1} \gamma_i = 1$: Stop Calibration

2.3 HL-RDHM Calibration: Implementation Details

In this study, the target distribution is the full posterior distribution $\pi(\boldsymbol{\theta}, \sigma_\delta^2, \sigma_\epsilon^2|\mathbf{Z})$ and the Bayesian hierarchical framework for the HL-RDHM distributed hydrological model calibra-

Algorithm 1: Fast Particle-based Calibration

Data: Z

Initialization:

Draw $\boldsymbol{\theta}_0^{(i)} \sim p(\boldsymbol{\theta})$ for particles $i = 1, ..., N$.

Set $w_0^{(i)} = 1/N$, $\gamma_0 = 0$, and K;

for *cycles* $t = 1, ..., T$ **do**

1. Compute full likelihood:

Calculate $L(\boldsymbol{\theta}_{t-1}^{(i)}|\mathbf{Z})$ for $i = 1, ..., N$;

2. Select optimal likelihood incorporation increment γ_t:

Set $\gamma_t = \text{argmin}_\gamma\{(ESS_{\gamma_t} - ESS_{thresh})^2\}$, where $\gamma \in (0.1, 1 - \sum_{i=1}^{t-1} \gamma_{t-1})$

Note: $ESS_{\gamma_t} = \sum_{i=1}^{N} \frac{1}{w_t^{(i)2}}$ and $w_t^{(i)} \propto L(\boldsymbol{\theta}_t^{(i)}|\mathbf{Z})^{\gamma_t}$;

3. Compute importance weights:

$w_t^{(i)} \propto w_{t-1}^{(i)} \times L(\boldsymbol{\theta}_t^{(i)}|\mathbf{Z})^{\gamma_t}$;

4. Re-sample particles:

Draw $\boldsymbol{\theta}_t^{(i)}$ from $\{\boldsymbol{\theta}_{t-1}^{(1)}, ..., \boldsymbol{\theta}_{t-1}^{(N)}\}$ with probabilities $\propto \{w_t^{(1)}, ..., w_t^{(N)}\}$;

5. Set intermediate posterior distribution:

Set $\pi_t(\boldsymbol{\theta}|\mathbf{Z}) \propto L(\boldsymbol{\theta}_i|\mathbf{Z})^{\tilde{\gamma}}\pi(\boldsymbol{\theta})$, where $\tilde{\gamma} = \sum_{j=1}^{t} \gamma_j$;

6. Mutation:

Using each particle $(\boldsymbol{\theta}_t^{(1)}, ..., \boldsymbol{\theta}_t^{(N)})$ as the initial value, run N chains of an MCMC algorithm with target distribution $\pi_t(\boldsymbol{\theta}|Z)$ for $2K$ iterations

7. Check stopping criterion:

Compute $\delta_B = D_B(h(\boldsymbol{\theta}_t^K), h(\boldsymbol{\theta}_t^{2K}))$;

if $\delta_B < \epsilon_B$ **then**

Set $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}_t^{(i),2K}$;

else

Run K additional updates and re-evaluate stopping criterion

Continue until stopping criterion is met

8. Stop when full likelihood is incorporated;

if $\sum_{i=1}^{N} \gamma_t = 1$ **then**

End Algorithm;

else

Reset weights: $w_t^{(i)} = 1/N$ for particles $i = 1, ..., N$;

Set t=t+1 and return to Step 1;

tion is as follows:

$$\text{Data Model:} \quad \mathbf{Z}|\mathbf{Y}(\cdot), \boldsymbol{\theta}, \boldsymbol{\delta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{Y}(\boldsymbol{\theta}) + \boldsymbol{\delta}, \sigma_\epsilon^2 \mathcal{I}) \tag{A8}$$

$$\text{Process Model:} \quad \boldsymbol{\delta}|\sigma_\delta^2 \sim \mathcal{N}(\mathbf{0}, \sigma_\delta^2 \mathcal{I}) \tag{A9}$$

$$\text{Parameter Model:} \quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \quad \sigma_\delta^2 \sim p(\sigma_\delta^2), \quad \sigma_\epsilon^2 \sim p(\sigma_\epsilon^2) \tag{A10}$$

where $p(\boldsymbol{\theta})$, $p(\sigma_\delta^2)$, and $p(\sigma_\epsilon^2)$ denote the prior distributions of $\boldsymbol{\theta}$, σ_δ^2, and σ_ϵ^2, respectively. For $p(\boldsymbol{\theta})$, we place a priori independent uniform priors on each of the model parameters with ranges (lower and upper bounds) based on domain-area expertise.

Instead of estimating the nuisance parameters σ_δ^2 and σ_ϵ^2 separately, we chose to combine these as $\sigma^2 = \sigma_\delta^2 + \sigma_\epsilon^2$. We place a standard non-informative inverse gamma prior on the combined error variance $\sigma_\epsilon^2 \sim IG(0.2, 0.2)$. The updated Bayesian hierarchical framework is:

$$\text{Data Model:} \quad \mathbf{Z}|\mathbf{Y}(\cdot), \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\mathbf{Y}(\boldsymbol{\theta}), \sigma^2 \mathcal{I}) \tag{A11}$$

$$\text{Parameter Model:} \quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \quad \sigma^2 \sim p(\sigma^2) \tag{A12}$$

While much of the fast particle-based approach is automated, the user must select the: (1) total number of particles, N; (2) baseline number of Metropolis-Hastings updates run before checking the stopping criterion, K; (3) minimum incorporation γ_{min} at each cycle; and (4) the effective sample size threshold ESS_{thresh}. We chose $N = 2015$ particles based on the available resources. On the Cheyenne HPC, this requires 56 nodes with 36 processors per node. For the stopping criterion, we use $k = 7$ as the baseline length. The floor for the incorporation increment is fixed at $\gamma_{min} = 0.1$ such that we incoporate at least $L(\boldsymbol{\theta}|\mathbf{Z})^{0.1}$ into the intermediate posterior at each cycle. Finally, the $ESS_{thresh} = N/2$, which is the typical threshold that activates resampling in many sequential Monte Carlo methods (Del Moral et al., 2006).We calibrate the HL-RDHM distributed hydrological model using Cheyenne (Computational and Information Systems Laboratory, 2017), a 5.34-petaflops high performance computer operated by the National Center for Atmospheric Research (NCAR). We employ message passing interface (MPI) and the `R` package `Rmpi` for any parallelized operations such as computing importance weights and particle mutation.

The prior distribution $p(\boldsymbol{\theta}_j)$ for the j-th HL-RDHM model parameters follow a univariate uniform distribution with lower and upper bounds specified by our hydrological model experts. $\boldsymbol{\theta}_j \sim Unif(l_j, u_j)$ with hyperparameters l_j(lower bound) and u_j (upper bound) specified in Table S1. We place a standard non-informative inverse gamma prior on the combined error variance $\sigma^2 \sim IG(\alpha_{\sigma^2}, \beta_{\sigma^2})$ where $\alpha_{\sigma^2} = 0.2$ and $\beta_{\sigma^2} = 0.2$.

References

Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., et al. (2017). Global projections of river flood risk in a warmer world. *Earth's Future*. https://doi.org/10.1002/2016ef000485

Anderson, R. M., Koren, V. I., & Reed, S. M. (2006). Using SSURGO data to improve Sacramento Model a priori parameter estimates. *Journal of Hydrology*, *320*(1), 103–116.

Asher, M. J., Croke, B. F. W., Jakeman, A. J., & Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research*. https://doi.org/10.1002/2015wr016967

Ashley, S. T., & Ashley, W. S. (2008). Flood Fatalities in the United States. *Journal of Applied Meteorology and Climatology*, *47*(3), 805–818.

Bain, A., & Crisan, D. (2008). *Fundamentals of Stochastic Filtering*. Springer Science & Business Media.

Bates, P. D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., et al. (2021). Combined modeling of US fluvial, pluvial, and coastal flood hazard under current and future climates. *Water Resources Research*, *57*(2). https://doi.org/10.1029/2020wr028673

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., et al. (2007a). A Framework for Validation of Computer Models. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *49*(2), 138–154.

Bayarri, M. J., Walsh, D., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., et al. (2007b). Computer model validation with functional output. *The Annals of Statistics*, *35*(5), 1874–1906.

Beven, K. (2014). The GLUE Methodology for Model Calibration with Uncertainty. *Applied Uncertainty Analysis for Flood Risk Management*. https://doi.org/10.1142/9781848162716_0006

Bhat, K. S., Haran, M., Goes, M., & Chen, M. (2010). Computer model calibration with multivariate spatial output: A case study. *Frontiers of Statistical Decision Making and Bayesian Analysis*, 168–184.

Bhattacharyya, A. (1946). On a Measure of Divergence between Two Multinomial Populations. *Journal of the Indian Society of Agricultural Statistics. Indian Society of Agricultural Statistics*, *7*(4), 401–406.

Bitew, M. M., & Gebremichael, M. (2011). Evaluation of satellite rainfall products through hydrologic simulation in a fully distributed hydrologic model. *Water Resources Research*, *47*(6). https://doi.org/10.1029/2010wr009917

Boulange, J., Hanasaki, N., Yamazaki, D., & Pokhrel, Y. (2021). Role of dams in reducing global flood exposure under climate change. *Nature Communications*, *12*(1), 417.

Bowman, A. L., Franz, K. J., & Hogue, T. S. (2017). Case Studies of a MODIS-Based Potential Evapotranspiration Input to the Sacramento Soil Moisture Accounting Model. *Journal of Hydrometeorology*, *18*(1), 151–158.

Braak, C. J. F. T. (2006). A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249.

Brunner, G. W. (1995). *HEC-RAS River Analysis System. Hydraulic Reference Manual. Version 1.0*. Hydrologic Engineering Center Davis CA. Retrieved from

https://apps.dtic.mil/sti/citations/ADA311952

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, *30*(11), 114007.

Carlberg, B., Franz, K., & Gallus, W. (2020). A Method to Account for QPF Spatial Displacement Errors in Short-Term Ensemble Streamflow Forecasting. *WATER*, *12*(12), 3505.

Chang, W., Haran, M., Olson, R., & Keller, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *The Annals of Applied Statistics*, *8*(2), 649–673.

Chang, W., Haran, M., Applegate, P., & Pollard, D. (2016). Calibrating an Ice Sheet Model Using High-Dimensional Binary Spatial Data. *Journal of the American Statistical Association*, *111*(513), 57–72.

Chester, M. V., Shane Underwood, B., & Samaras, C. (2020). Keeping infrastructure reliable under climate uncertainty. *Nature Climate Change*. https://doi.org/10.1038/s41558-020-0741-0

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*(3), 539–552.

Computational and Information Systems Laboratory (2017). Cheyenne: HPE/SGI ICE XA System (University Community Computing). Boulder, CO: National Center for Atmospheric Research. doi:10.5065/D6RX99HX.

Constantine, P. G., Dow, E., & Wang, Q. (2014). Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces. *SIAM Journal on Scientific Computing*. https://doi.org/10.1137/130916138

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

Craig, P. S., Goldstein, M., Seheult, A. H., & Smith, J. A. (1997). Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments. In *Case Studies in Bayesian Statistics* (pp. 37–93). Springer New York.

Crisan, D., & Doucet, A. (2000). Convergence of sequential Monte Carlo methods. *Signal Processing Group, Department of Engineering, University of Cambridge, Technical Report CUEDIF-INFENGrrR38*, *1*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.361.3193&rep=rep1&type=pdf

Davis and Skaggs. (1992). *Catalog of Residential Depth-Damage Functions Used by the Army Corps of Engineers in Flood Damage Estimation*. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a255462.pdf

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. Journal of the Royal Statistical Society: Series B, 68(3):411– 436.

Didier, D., Baudry, J., Bernatchez, P., Dumont, D., Sadegh, M., Bismuth, E., et al. (2019). Multihazard simulation for coastal flood mapping: Bathtub versus numerical modelling in an open estuary, Eastern Canada. *Journal of Flood Risk Management*, *12*(S1), e12505.

Doucet, A., Godsill, S., & Andrieu, C. (2000). *Statistics and Computing*, *10*(3), 197–208.

Doucet, A., de Freitas, N., & Gordon, N. (2001). An Introduction to Sequential Monte Carlo Methods. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice* (pp. 3–14). New York, NY: Springer New York.

Edwards, N. R., Cameron, D., & Rougier, J. (2011). Precalibrating an intermediate complexity climate

model. *Climate Dynamics*, *37*(7-8), 1469–1482.

Fares, A., Awal, R., Michaud, J., Chu, P.-S., Fares, S., Kodama, K., & Rosener, M. (2014). Rainfall-runoff modeling in a flashy tropical watershed using the distributed HL-RDHM model. *Journal of Hydrology*, *519*, 3436–3447.

FEMA, 2019: Flood Insurance Rate Map (FIRM). Federal Emergency Management Agency, https://www.fema.gov/flood-insurance-ratemap-firm.

Fereshtehpour, M., & Karamouz, M. (2018). DEM resolution effects on coastal flood vulnerability assessment: Deterministic and probabilistic approach. *Water Resources Research*, *54*(7), 4965–4982.

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4). https://doi.org/10.1029/2018ms001453

Gilks, W. R., & Berzuini, C. (2001). Following a moving target-Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *63*(1), 127–146.

Gomez, M., Sharma, S., Reed, S., & Mejia, A. (2019). Skill of ensemble flood inundation forecasts at short- to medium-range timescales. *Journal of Hydrology*, *568*, 207–220.

Goodman, J., & Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, *5*(1), 65–80.

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-radar and signal processing* (Vol. 140, pp. 107–113). IET.

Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., et al. (2020). Sensitivity Analysis-Based Automatic Parameter Calibration of the VIC Model for Streamflow Simulations Over China. *Water Resources Research*. https://doi.org/10.1029/2019wr025968

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC.

Gramacy, R. B., & Apley, D. W. (2015). Local Gaussian Process Approximation for Large Computer Experiments. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *24*(2), 561–578.

Helton, J. C., & Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, *81*(1), 23–69.

Herman, J. D., Reed, P. M., & Wagener, T. (2013). Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resources Research*, *49*(3), 1400–1414.

Higdon, D. (2003). for inference in computationally intensive inverse problems. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting* (p. 181). Oxford University Press.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., & Ryne, R. D. (2004). Combining Field Data and Computer Simulations for Calibration and Prediction. *SIAM Journal of Scientific Computing*, *26*(2), 448–466.

Higdon, D., Gattiker, J., Williams, B., & Rightley, M. (2008). Computer Model Calibration Using High-Dimensional Output. *Journal of the American Statistical Association*, *103*(482), 570–583.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., & Wilkinson, R. D. (2010). A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. *Climate Dynamics*, *35*(5), 785–806.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., et al. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States--representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, *81*(5), 345–354.

Hsu, K.-L., Moradkhani, H., & Sorooshian, S. (2009). A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2008wr006824

Hu, J., Chen, S., Behrangi, A., & Yuan, H. (2019). Parametric uncertainty assessment in hydrological modeling using the generalized polynomial chaos expansion. *Journal of Hydrology*, *579*, 124158.

Hwang, J. T., & Martins, J. R. R. A. (2018). A fast-prediction surrogate model for large datasets. *Aerospace Science and Technology*, *75*, 74–87.

Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., & Sharma, A. (2011). Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and sequential Monte Carlo samplers. *Water Resources Research*, *47*(7). https://doi.org/10.1029/2010wr010217

Judi, D. R., Rakowski, C. L., Waichler, S. R., Feng, Y., & Wigmosta, M. S. (2018). Integrated Modeling Approach for the Development of Climate-Informed, Actionable Information. *WATER*, *10*(6), 775.

Kalyanaraman, J., Kawajiri, Y., Lively, R. P., & Realff, M. J. (2016). Uncertainty quantification via bayesian inference using sequential monte carlo methods for CO2adsorption process. *AIChE Journal*. https://doi.org/10.1002/aic.15381

Kamali, B., Mousavi, S. J., & Abbaspour, K. C. (2013). Automatic calibration of HEC-HMS using single-objective and multi-objective PSO algorithms. *Hydrological Processes*, *27*(26), 4028–4042.

Kantas, N., Beskos, A., & Jasra, A. (2014). Sequential Monte Carlo Methods for High-Dimensional Inverse Problems: A Case Study for the Navier--Stokes Equations. *SIAM/ASA Journal on Uncertainty Quantification*, *2*(1), 464–489.

Kavetski, D., Fenicia, F., Reichert, P., & Albert, C. (2018). Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications. *Water Resources Research*. https://doi.org/10.1002/2017wr020528

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *63*(3), 425–464.

Koren, V., Reed, S., Smith, M., Zhang, Z., & Seo, D.-J. (2004). Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. *Journal of Hydrology*, *291*(3), 297–318.

Kuzmin, V., Seo, D.-J., & Koren, V. (2008). Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search. *Journal of Hydrology*, *353*(1), 109–128.

Lahmers, T. M., Hazenberg, P., Gupta, H., Castro, C., Gochis, D., Dugger, A., et al. (2021). Evaluation of NOAA National Water Model Parameter Calibration in Semiarid Environments Prone to Channel Infiltration. *Journal of Hydrometeorology*, *22*(11), 2939–2969.

Lataniotis, C., Marelli, S., & Sudret, B. (2020). EXTENDING CLASSICAL SURROGATE MODELING TO HIGH DIMENSIONS THROUGH SUPERVISED DIMENSIONALITY REDUCTION: A DATA-DRIVEN APPROACH. *International Journal for Uncertainty Quantification*, *10*(1).

https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020031935

Lee, B. S., Haran, M., Fuller, R. W., Pollard, D., & Keller, K. (2020). A fast particle-based approach for calibrating a 3-D model of the Antarctic ice sheet. *The Annals of Applied Statistics*, *14*(2), 605–634.

Lempert, R. J. (2002). A new decision sciences for complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, *99 Suppl 3*, 7309–7313.

Li, T., Sun, S., Sattar, T. P., & Corchado, J. M. (2014). Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with Applications*, *41*(8), 3944–3954.

Liang, F., & Wong, W. H. (2001). Real-Parameter Evolutionary Monte Carlo With Applications to Bayesian Mixture Models. *Journal of the American Statistical Association*, *96*(454), 653–666.

Liu, J., & West, M. (2001). Combined Parameter and State Estimation in Simulation-Based Filtering. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice* (pp. 197–223). New York, NY: Springer New York.

Liu, J. S., Liang, F., & Wong, W. H. (2000). The Multiple-Try Method and Local Optimization in Metropolis Sampling. *Journal of the American Statistical Association*, *95*(449), 121–134.

Liu, X., & Guillas, S. (2017). Dimension Reduction for Gaussian Process Emulation: An Application to the Influence of Bathymetry on Tsunami Heights. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), 787–812.

Liu, Y., Hejazi, M., Li, H., Zhang, X., & Leng, G. (2018). A hydrological emulator for global applications – HE v1.0.0. *Geoscientific Model Development*. https://doi.org/10.5194/gmd-11-1077-2018

Liu, Z., & Merwade, V. (2018). Accounting for model structure, parameter and input forcing uncertainty in flood inundation modeling using Bayesian model averaging. *Journal of Hydrology*, *565*, 138–149.

Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., et al. (2018). An Efficient Surrogate Model for Emulation and Physics Extraction of Large Eddy Simulations. *Journal of the American Statistical Association*, *113*(524), 1443–1456.

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145–2166.

Mckay, M. D., Beckman, R. J., & Conover, W. J. (2000). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *42*(1), 55–61.

McEnery, J., Ingram, J., Duan, Q., Adams, T., & Anderson, L. (2005). NOAA'S ADVANCED HYDROLOGIC PREDICTION SERVICE: Building Pathways for Better Science in Water Forecasting. *Bulletin of the American Meteorological Society*, *86*(3), 375–386.

Mejia, A. I., & Reed, S. M. (2011). Evaluating the effects of parameterized cross section shapes and simplified routing with a coupled distributed hydrologic and hydraulic model. *Journal of Hydrology*, *409*(1), 512–524.

Mendoza, P. A., Clark, M. P., Mizukami, N., Newman, A. J., Barlage, M., Gutmann, E. D., et al. (2015). Effects of Hydrologic Model Choice and Calibration on the Portrayal of Climate Change Impacts. *Journal of Hydrometeorology*, *16*(2), 762–780.

Merz, B., Hall, J., Disse, M., & Schumann, A. (2010). Fluvial flood risk management in a changing world. *Natural Hazards and Earth System Sciences*, *10*(3), 509–527.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., & Kumar, R. (2019). On the choice of calibration metrics for "high-flow" estimation using hydrologic models. *Hydrology and Earth System Sciences*, *23*(6), 2601–2614.

Morzfeld, M., Day, M. S., Grout, R. W., Heng Pau, G. S., Finsterle, S. A., & Bell, J. B. (2018). Iterative Importance Sampling Algorithms for Parameter Estimation. *SIAM Journal of Scientific Computing*, *40*(2), B329–B352.

Murphy, A. H. (1970). THE RANKED PROBABILITY SCORE AND THE PROBABILITY SCORE: A COMPARISON. *Monthly Weather Review*, *98*(12), 917–924.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology and Climatology*, *12*(4), 595–600.

Murray, L. M., Lee, A., & Jacob, P. E. (2016). Parallel Resampling in the Particle Filter. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *25*(3), 789–805.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*(2), 125–139.

Neumann, T., & Ahrendt, K. (2013). Comparing The" Bathtub Method" With Mike 21 Hd Flow Model For Modelling Storm Surge Inundation. *Ecologic Institute, Berlin, Germany*. Retrieved from https://edoc.sub.uni-hamburg.de/klimawandel/frontdoor/deliver/index/docId/835/file/RADOST_BATHTUB_034.pdf

Oakley, J. E. (2009). Decision-Theoretic Sensitivity Analysis for Complex Computer Models. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *51*(2), 121–129.

Papaioannou, I., Papadimitriou, C., & Straub, D. (2016). Sequential importance sampling for structural reliability analysis. *Structural Safety*, *62*, 66–75.

Park, S., Hwang, J. P., Kim, E., & Kang, H.-J. (2009). A New Evolutionary Particle Filter for the Prevention of Sample Impoverishment. *IEEE Transactions on Evolutionary Computation*, *13*(4), 801–809.

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

Prat, O. P., & Nelson, B. R. (2015). Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-19-2037-2015

Rafieeinasab, A., Norouzi, A., Kim, S., Habibi, H., Nazari, B., Seo, D.-J., et al. (2015). Toward high-resolution flash flood prediction in large urban areas – Analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling. *Journal of Hydrology* , *531*, 370–388.

Raje, D., & Krishnan, R. (2012). Bayesian parameter uncertainty modeling in a macroscale hydrologic model and its impact on Indian river basin hydrology under climate change. *Water Resources Research*, *48*(8). https://doi.org/10.1029/2011wr011123

Rajib, A., Liu, Z., Merwade, V., Tavakoly, A. A., & Follum, M. L. (2020). Towards a large-scale locally relevant flood inundation modeling framework using SWAT and LISFLOOD-FP. *Journal of*

Hydrology, *581*, 124406.

Razavi, S., & Tolson, B. A. (2013). An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, *49*(12), 8418–8431.

Read, L. K., & Vogel, R. M. (2015). Reliability, return periods, and risk under nonstationarity. *Water Resources Research*, *51*(8), 6381–6398.

Reed, S., Schaake, J., & Zhang, Z. (2007). A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrology*, *337*(3), 402–420.

Rojas, M., Quintero, F., & Krajewski, W. F. (2020). Performance of the national water model in Iowa using independent observations. *Journal of the American Water Resources Association*, *56*(4), 568–585.

Ruckert, K. L., Srikrishnan, V., & Keller, K. (2019). Characterizing the deep uncertainties surrounding coastal flood hazard projections: A case study for Norfolk, VA. *Scientific Reports*, *9*(1), 11373.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Schweizerische Monatsschrift Fur Zahnheilkunde = Revue Mensuelle Suisse D'odonto-Stomatologie / SSO*, *4*(4), 409–423.

Salas, F. R., Somos-Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., et al. (2018a). Towards real-time continental scale streamflow simulation in continuous and discrete space. *Journal of the American Water Resources Association*, *54*(1), 7–27.

Salas, J. D., Obeysekera, J., & Vogel, R. M. (2018b). Techniques for assessing water infrastructure for nonstationary extreme events: a review. *Hydrological Sciences Journal*, *63*(3), 325–352.

Sanders, B. F., Schubert, J. E., Goodrich, K. A., Houston, D., Feldman, D. L., Basolo, V., et al. (2020). Collaborative Modeling With Fine-Resolution Data Enhances Flood Awareness, Minimizes Differences in Flood Perception, and Produces Actionable Flood Maps. *Earth's Future*. https://doi.org/10.1029/2019ef001391

Scawthorn, C., Blais, N., Seligson, H., Tate, E., Mifflin, E., Thomas, W., et al. (2006). HAZUS-MH flood loss estimation methodology. I: Overview and flood hazard characterization. *Natural Hazards Review*, *7*(2), 60–71.

Shafii, M., Tolson, B., & Shawn Matott, L. (2015). Addressing subjective decision-making inherent in GLUE-based multi-criteria rainfall–runoff model calibration. *Journal of Hydrology*, *523*, 693–705.

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological model diversity enhances streamflow forecast skill at short- to medium-range timescales. *Water Resources Research*, *55*(2), 1510–1530.

Sharma, S., Gomez, M., Keller, K., Nicholas, R., & Mejia, A. (2021). Regional Flood Risk Projections under Climate Change. *Journal of Hydrometeorology*, *-1*(aop). https://doi.org/10.1175/JHM-D-20-0238.1

Siddique, R., & Mejia, A. (2017). Ensemble Streamflow Forecasting across the U.S. Mid-Atlantic Region with a Distributed Hydrological Model Forced by GEFS Reforecasts. *Journal of Hydrometeorology*, *18*(7), 1905–1928.

Stein, M. (1987). Large Sample Properties of Simulations Using Latin Hypercube Sampling. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *29*(2), 143–151.

Steinberg, D. M., & Lin, D. K. J. (2006). A construction method for orthogonal Latin hypercube designs.

Biometrika. https://doi.org/10.1093/biomet/93.2.279

Storn, R., & Price, K. (1997). *Journal of Global Optimization*, *11*(4), 341–359.

Su, Y., Feng, Q., Zhu, G., Gu, C., Wang, Y., Shang, S., et al. (2018). A hierarchical Bayesian approach for multi-site optimization of a satellite-based evapotranspiration model. *Hydrological Processes*, *32*(26), 3907–3923.

Tarawneh, E., Bridge, J., & Macdonald, N. (2016). A pre-calibration approach to select optimum inputs for hydrological models in data-scarce regions. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-20-4391-2016

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., et al. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, *596*(7870), 80–86.

Wasko, C., Westra, S., Nathan, R., Orr, H. G., Villarini, G., Villalobos Herrera, R., & Fowler, H. J. (2021). Incorporating climate change in flood estimation guidance. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *379*(2195), 20190548.

Wing, O. E. J., Bates, P. D., Smith, A. M., Sampson, C. C., Johnson, K. A., Fargione, J., & Morefield, P. (2018). Estimates of present and future flood risk in the conterminous United States. *Environmental Research Letters: ERL [Web Site]*, *13*(3), 034023.

Wing, O. E. J., Pinter, N., Bates, P. D., & Kousky, C. (2020). New insights into US flood vulnerability revealed from flood insurance big data. *Nature Communications*, *11*(1), 1444.

Winsemius, H. C., Jeroen C J, van Beek, L. P. H., Bierkens, M. F. P., Bouwman, A., Jongman, B., et al. (2015). Global drivers of future river flood risk. *Nature Climate Change*, *6*(4), 381–385.

Wong, T. E., & Keller, K. (2017). Deep Uncertainty Surrounding Coastal Flood Risk Projections: A Case Study for New Orleans. *Earth's Future*. https://doi.org/10.1002/2017ef000607

Yunus, A. P., Avtar, R., Kraines, S., Yamamuro, M., Lindberg, F., & Grimmond, C. S. B. (2016). Uncertainties in Tidally Adjusted Estimates of Sea Level Rise Flooding (Bathtub Model) for the Greater London. *Remote Sensing*, *8*(5), 366.

Zarekarizi, M., Srikrishnan, V., & Keller, K. (2020). Neglecting uncertainties biases house-elevation decisions to manage riverine flood risks. *Nature Communications*, *11*(1), 5361.

Zarzar, C. M., Hosseiny, H., Siddique, R., Gomez, M., Smith, V., Mejia, A., & Dyer, J. (2018). A hydraulic MultiModel ensemble framework for visualizing flood inundation uncertainty. *Journal of the American Water Resources Association*, *54*(4), 807–819.

Zhu, G., Li, X., Ma, J., Wang, Y., Liu, S., Huang, C., et al. (2018). A new moving strategy for the sequential Monte Carlo approach in optimizing the hydrological model parameters. *Advances in Water Resources*, *114*, 164–179.