**Accounting model parametric uncertainty improves flood risk estimates**

Sanjib Sharma1\*, Benjamin Seiyon Lee2, Iman Hosseini-Shakib1, Murali Haran3, Klaus Keller1,4,5

1Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA, USA

2Department of Statistics, George Mason University, Fairfax, VA, USA

3Department of Statistics, Pennsylvania State University, University Park, PA, USA

4Department of Geosciences, Pennsylvania State University, University Park, PA, USA

5Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

\*Corresponding author (sanjibsharma66@gmail.com)

Projected Submission on Water Resources Research

**Key points**

● Current approaches to characterize flood hazards often sample only a relatively small subset of the known unknowns such as hydrologic model parameters.

● We implement a sequential Monte Carlo particle-based approach to fully characterize the hydromodel parametric uncertainty.

● Accounting model parametric uncertainty improves flood hazards and risk predictions.

**Keywords**

Hydromodel calibration, Particle-based approach,

Uncertainty characterization, Flood hazards, Flood damage

**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 such as model parameters. This approach neglects the impacts of key uncertainties on hazards and dynamics. Here we develop a Bayesian data-model fusion framework for calibrating computationally expensive distributed hydrologic models. We compare three different calibration approaches: (1) stepwise line search; (2) precalibration or screening; and (3) the Fast Model Calibrations (FaMoS) approach. Specifically, we deploy a particle-based approach that takes advantage of the massive parallelization afforded by modern high-performance computing systems. Results show that using a best estimate parameter set drastically underestimate the extreme flood events. We find that considering model parametric uncertainty improves flood hazard and damage estimates.

**1. Motivation and Introduction**

Floods pose major risks to people and property [(Alfieri et al., 2017; Wing et al., 2018; Winsemius et al., 2015)](https://paperpile.com/c/9lxFNc/GrqrC+qyH3w+fQXtd). These risks are dynamic and deeply uncertain. 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 [(Chester et al., 2020; Liu & Merwade, 2018; Salas et al., 2018; Wasko et al., 2021; Wong & Keller, 2017)](https://paperpile.com/c/9lxFNc/8gr4l+9gFVx+XijHS+72IVp+nUc7o).

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 predictions are subject to deep uncertainties. One key deep uncertainty that is often overlooked is uncertainty in the model parameter (parametric uncertainty). Parametric uncertainty stems from the nature of unknown model parameters as well as difficulties associated with calibrating the model parameters and the divergent expert assessments. 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 limited observability. In these cases, 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.

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)](https://paperpile.com/c/9lxFNc/u3Xxf+JLmo8+qyH3w+oByr2+1bLwF+5VPxv). 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 loosely coupled land surface hydrologic and river hydraulic modeling framework to provide regional flood hazard and risk estimates. Recently, Bates et al. (2021) presented current and future flood risk estimates for all properties in the conterminous United States using combined modeling of river, coastal, or rainfall flooding. These studies typically rely on obtaining an optimal parameter set that produces the best possible agreement between simulated and observed streamflow hydrographs at target locations. Previous studies break important new ground, but neglect the impacts of key parametric uncertainties on hazards and dynamics. Neglecting parametric uncertainties can underestimate the tails of flood hazard probability distribution, and can result in poor decisions and outcomes [(Wong & Keller, 2017; Zarekarizi et al., 2020)](https://paperpile.com/c/9lxFNc/rVyIU+XijHS).

Hydrologic model calibration typically relies on manually adjusting a subset of model parameters (Bitew & Gebremichael, 2011; Siddique & Mejia, 2017), automatic parameter optimization [(Kamali et al., 2013; Razavi & Tolson, 2013; Van Liew et al., 2005)](https://paperpile.com/c/9lxFNc/mIBQy+aRP7D+4ELbE), and/or surrogate methods such as Gaussian process-based emulators to provide best-fit parameter sets (Gou et al., 2020; Pianosi et al., 2016; [(Razavi & Tolson, 2013)](https://paperpile.com/c/9lxFNc/aRP7D)​​. Manual calibration relies on visual inspection of streamflow hydrograph and a trial and error-based procedure; hence, this method can be increasingly labor-intensive and time-consuming (Siddique & Mejia, 2017). 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). Recently, Gou et al. (2020) developed 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 fewer model parameters because training a surrogate model can be computationally prohibitive in high-dimensional settings. Hence, surrogate-based approaches typically neglect the high dimensionality of hydrologic models or are computationally impractical because they require too many model runs and/or too much computational resources. Therefore, surrogate-based methods may be inappropriate for estimating a large set of distributed parameters in fine-scale hydrologic modeling.

Bayesian calibration of hydrologic models have become increasingly popular (Jeremiah et al., 2011; Shafii et al., 2015; Su et al., 2018; Zhu et al., 2018; [Hsu et al., 2009; Kavetski et al., 2018; Raje & Krishnan, 2012; Razavi & Tolson, 2013)](https://paperpile.com/c/9lxFNc/aRP7D+E8cn+SiyT+48nY). 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 (PT-JPL) model using observed evapotranspiration measurements. Given the shorter single model run times, the hierarchical model is fit using the Differential Evolution Markov Chain, a population MCMC algorithm. Zhu et al. (2018) calibrates eight parameters-conceptual water balance model using a Particle Evolution Metropolis Sequential Monte Carlo (PEM-SMC). The PEM-SMC algorithm evaluates the water balance model 2000 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) lumped hydrological model; (2) limited model parameters; (3) low-to-moderate flow threshold; and (4) relatively small basin. However, the computational requirement can be exponentially higher in the context of fully distributed hydrological modeling over the large basin and with a large number of sensitive parameters.

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

**2. Bayesian Model Calibration**

In model calibration, the objective is two-fold: (1) infer the input parameters (i.e. point estimates); and (2) quantify the uncertainty underlying the estimated input parameters (i.e. interval estimates). Other sources of uncertainty such as model-observation discrepancy (Kennedy and O’Hagan, 2001; Bayarri et al., 2007; Brynjarsdottir and O’Hagan, 2014) and measurement errors may directly influence parameter estimation. The Bayesian model calibration framework (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 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. The posterior distribution **)** is defined as:

**,**

where denotes the model parameters, is the variance of the independent and identically distributed observational error, and is the discrepancy term.

For complex deterministic models, the posterior distribution may not be available in the closed form (Higdon et al., 2003; Kaipo et al., 2000); hence, we must approximate the posterior via sampling approaches such as Markov chain Monte Carlo (MCMC) or Sequential Monte Carlo. For the sampling approaches, there exists two major constraints: (1) walltimes for a single model evaluation; and (2) the number of model parameters to be calibrated. Markov chain Monte Carlo methods with the true model are sensible for models with short single model run times (less than three seconds) and many model parameters. Emulation-calibration approaches replace the hydrologic model with a faster surrogate model, or emulator, and then sample from the posterior distribution via MCMC. However, emulation-calibration methods tend to be limited to hydrologic models with few model parameters. Sequential Monte Carlo methods (SMC) (Lee et al. 2021, GA Tech paper, other SMC paper) provides a practical way to calibrate high dimensional models with a larger number of 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 are defined as:

(1)

where is the target function and is the important function. In this context, we specify the target function as the posterior distribution of the model parameters and importance function as the prior distribution of the parameters . We approximate the posterior distribution using the weighted empirical distribution **)** defined as:

(2)

where is the importance weight and is a Dirac measure at 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. Suppose there exists 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:

**,**

where denotes the incorporation factor such that . Note that the 0-th intermediate posterior distribution () is simply the prior distribution with incorporation factor . Likewise, the T-th intermediate posterior distribution () is the full posterior distribution since 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 must “jitter” or “mutate” the particles through a carefully constructed kernel function (Gilks and Berzuini, 2001; Liu and West, 2001; Li et al., 2013). 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:

(3)

Where is the i-th mutated particle, are the corresponding weights from the T-th cycle, and is a Dirac measure at .

**3. Experimental Design**

We focus on the West Branch Susquehanna River basin in the United States Middle Atlantic region as flooding is a regional concern. This region has a relatively high frequency of extreme weather events, making it particularly vulnerable to damaging flood events. The most destructive floods in the Susquehanna River basins 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 use the National Oceanic and Atmospheric Administration's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) (Koren et al., 2004). Within HL-RDHM, we deploy the Sacramento Soil Moisture Accounting model with Heat Transfer (SAC-HT) to represent hillslope runoff generation, and the SNOW-17 module to represent snow accumulation and melting. 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. We run HL-RDHM in a fully distributed mode at a spatial resolution of 2 km. The 2 × 2 km2 resolution mainly allows for a more realistic representation of the stream network.

We use three main datasets: multisensor precipitation estimates, gridded near-surface air temperature, and streamflow. We use multisensor precipitation estimates and gridded near-surface air temperature to run the hydrological model for parameter calibration purposes and to initialize the model. Multisensor precipitation estimates represents a continuous time series of hourly, gridded precipitation observations at 4 × 4 km2 cells, which are produced by combining multiple radar estimates and rain gauge measurements. The gridded near-surface air temperature data at 4 × 4 km2 resolution were developed by combining multiple temperature observation networks. We obtain streamflow observation at basin outlet from the United States Geological Survey (USGS 01554000). The selected gauge station represents the drainage area of 47, 396 km2.

We calibrate the model for the period of 2004-2009 and use 2010-2012 for evaluation. We use the year 2003 to spin up the model. As part of the calibration process, we select twelve model parameters associated with each model grid cell (Table 1). The sensitive parameters are associated with different hydrodynamic processes related to baseflow, percolation, evaporation, storm runoff, channel routing and snow processes (Supplementary Figure S1).

**3. Results**

We first generate streamflow simulations using the "best” parameter estimates obtained via the stepwise line search approach (Figure 2). The 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. Stepwise line search can deliberately avoid lower probability outcomes and excludes comprehensive sampling of the parametric distribution. However, the stepwise line search approach substantially underestimates the high streamflow (Figure 2).

​​We account for parametric uncertainty using precalibration and FaMoS (Supplemental Figure S1). Characterizing parametric uncertainty requires knowledge of model behavior throughout a typically high-dimensional parameter space. This requires large ensembles of model runs. 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. Precalibration provides a relatively low-cost method to explore the high-dimensional parameter space and extract as much information as possible from the limited number of 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. We select an ensemble of 165 runs that fall within the 25% -175% of the observed streamflow. Note that specifying these bounds can be arbitrary, yet this choice directly impacts the resulting collection of model runs. 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. In any of these cases, it would imply a more controversial a priori decision. In addition, we do not want to overlook the upper tails of flood hazards, which are often associated with high-risk and high-cost events.

FaMoS characterizes and quantifies the parametric uncertainty by updating our prior knowledge of the model parameters in light of new observations. 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. 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 (Supplemental 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 locates a “target” region that contain the most plausible sets of model parameters. In contrast, precalibration attempts to locate this “target” region using a single initial ensemble of model runs.

We evaluate the calibrated model performance using several decision-relevant metrics. We use traditional deterministic metrics such as the Klinge-Gupta Efficiency [(Mizukami et al., 2019)](https://paperpile.com/c/9lxFNc/NT0K), 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 [(Murphy, 1973)](https://paperpile.com/c/9lxFNc/XLdw) and the Continuous Ranked Probability Score [(Murphy, 1970)](https://paperpile.com/c/9lxFNc/xrOj). 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 evaluation is focused on high flows by choosing the river flow that exceeds NOAA’s Action Stage [(McEnery et al., 2005)](https://paperpile.com/c/9lxFNc/yPBP). 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 type of mitigation action in preparation for possible significant hydrologic activity.

Accounting for parametric uncertainty improves model performance metrics for the calibration data and out-of-sample predictions (Figure 3). We compute the skill score with reference to raw (uncalibrated) model runs using default parameter estimates obtained from several previous studies (Anderson et al. 2006, Reed et al. 2004). 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.

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 approach 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. For example, consider Tropical Storm Lee with streamflow observation of 11292 m3/sec. Precalibration provides a flood peak prediction of 10539 m3/sec and prediction interval (5%-95% credible interval) range from 6359 m3/sec to 14222 m3/sec (width=7863 m3/sec). FaMoS has a corresponding flood peak prediction of 11467 m3/sec with a credible interval ranging from 9925 m3/sec to 13121 m3/sec (width=3196 m3/sec).

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). In the context of flood risk management, decision makers must choose between two alternatives on the basis of a probability prediction for an event, with one of the decisions being preferred if the event does not occur, and the other being preferred if the event does occur. Thus, it is critical to have a prediction system that correctly predicts the occurrence of an event (probability of detection) while avoiding too many incorrect predictions when it does not occur (probability of false detection). We examine the relative operating characteristics (ROC) score based on the area underneath the ROC curve. The ROC score assess 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 [(Mason & Graham, 2002)](https://paperpile.com/c/9lxFNc/uUnA) (Figure 5). Streamflow predictions obtained using the FaMoS parameter distribution exhibit better discriminatory ability than the stepwise line search and precalibration. Stepwise line search shows a relatively poor ability to discriminate between different events. This can lead to bad decisions and outcomes related to flood risk management, for example,

Neglecting parametric uncertainty also underestimates potential flood damage (Figure 6). To quantify damage from flood hazards, we consider 2,000 hypothetical houses to quantify the damage from flood hazards. We assess damage for a certain depth of water in a house by using a Bathtub terrain flood model and a vulnerability model. 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. A common source of vulnerability model (depth-damage function) in damage assessment studies in the U.S. is Hazard U.S. (HAZUS) provided by the Federal Emergency Management Agency (FEMA). We find that the stepwise line search tends to underestimate the flood damage. The underestimation bias increases as flood magnitude increases. Accounting 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 capture 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.

**4. Conclusions**

​​ We present a Bayesian data-model fusion framework for: (1) calibrating a distributed hydrologic model; and (2) demonstrating practical implications of neglecting key uncertainties on hazard and risk estimates. We calibrate a distributed hydrologic model using three different approaches - stepwise line search, precalibration, and FaMoS. Our analysis indicates that traditional approaches of model calibration using the best fit parameter set underestimate flood hazards. Precalibration improves flood hazards estimates over the best fit estimates, but provides a wider predictive intervals (i.e. highly uncertain estimates). Overall, FaMoS provide improved predictive skill than stepwise line search and precalibration. Our results demonstrate that neglecting model parametric uncertainty can substantially underestimate flood hazards and risk estimates.

We caution that 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 (Highdon et al., 2008; Chang et al. 2014) or eigenfunctions (Mak et al. 2020) 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 and 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 terrain model. Future work could use process-informed models to characterize the impacts of hydrodynamic processes in damage estimates. In addition, future work could sample the uncertainty surrounding the flood vulnerability of the building.

**5. Methods and Materials**

**5.1. Hydrometeorological observations**

We use three main observation datasets: multisensor precipitation estimates (MPEs), gridded near-surface air temperature, and daily streamflow. MPEs and gridded near-surface air temperature are used to run the hydrological model in simulation mode for parameter calibration purposes and to initialize the hydrological model. Both the MPEs and gridded near-surface air temperature data at 4 × 4 km2 resolution were provided by the NOAA's Middle Atlantic River Forecast Center. The MPEs represents a continuous time series of hourly, gridded precipitation observations at 4 × 4 km2 cells, which are produced by combining multiple radar estimates and rain gauge measurements. The gridded near-surface air temperature data at 4 × 4 km2 resolution were developed by combining multiple temperature observation networks. Daily streamflow observations for the selected basins were obtained from the United States Geological Survey (USGS 01554000).

**5.2. Distributed Hydrological Model**

We use the National Oceanic and Atmospheric Administration's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) (Koren et al., 2004). HL-RDHM is used as the spatially distributed hydrological model. The basin system is divided into regularly spaced, square grid cells to account for spatial heterogeneity. Each grid cell acts as a hillslope capable of generating surface, interflow, and groundwater runoff that discharges directly into the streams. The cells are connected to each other through the stream network system. Within HL-RDHM, we deploy the Sacramento Soil Moisture Accounting model with Heat Transfer (SAC-HT) to represent hillslope runoff generation, and the SNOW-17 module to represent snow accumulation and melting. 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. Likewise, flows in the stream network are routed downstream using a nonlinear kinematic wave algorithm that accounts for parameterized stream cross-section shapes. We run HL-RDHM in a fully distributed mode at a spatial resolution of 2 km. Further information about the HL-RDHM model can be found elsewhere (Koren et al., 2004; Reed et al., 2004; Anderson et al., 2006).

**5.3. Study Area**

We choose the West Branch Susquehanna River (WBSR) basin in the US Middle Atlantic region (MAR) as the study area. The WBSR is selected as flooding is an important regional concern. This region has a relatively high frequency of extreme weather events, making it particularly vulnerable to damaging flood events. The climate in the upper MAR, where the NBSR basin is located, can be classified as warm, humid summers and snowy, cold winters with frozen precipitation (Polsky et al., 2000). During the cool season, a positive North Atlantic Oscillation phase generally results in increased precipitation amounts and the occurrence of heavy snow (Durkee et al., 2007). Thus, flooding in the cool season is dominated by heavy precipitation events accompanied by snowmelt runoff. In the summer season, convective thunderstorms with increased intensity may lead to greater variability in streamflow. In the WBSR basin, we select the US Geological Survey daily gauge station (USGS 01554000) at Sunbury, Pennsylvania. The selected gauge station represents the drainage area of 47, 396 km2.

**5.4. Model Parameters**

​​As part of the calibration process, we select 12 out of the 17 model parameters associated with each model grid cell. Note that we adjust the parameter fields rather than the actual parameter values. The calibrated parameters are associated with baseflow, percolation, evaporation, storm runoff, and the channel routing process (Table 1). The most sensitive parameters are lower zone supplemental withdrawal rate (LZSK), upper zone free water maximum storage (UZFWM), Lower zone tension water maximum storage (LZTWM), Lower zone free water supplemental maximum storage (LZFSM), Lower zone free water primary maximum storage (LZFPM), Percolation equation exponent (REXP), Upper zone free water withdrawal rate (UZK), permanent impervious area (PCTIM), saturated impervious area (ADIMP), and the channel routing parameters (Figure 1). By comparing the performance of individual hydrographs during the cool season, we observe that high flow events are consistently underestimated with the default parameter set. To improve this, we adjust the SNOW-17 parameter, including the snow fall correction factor (SCF).

We specify univariate uniform prior distributions for all 12 model parameters. For each uniform prior distribution, we impose lower and upper bounds (i.e. the hyperparameters) specified by our hydrological model experts (Table 1). For the combined error variance parameter (), we designate an inverse gamma prior with hyperparameters scale and shape.

Table 1. HL-RDHM parameters and their sampled ranges for sensitivity analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/N** | **Parameter** | **Description** | **Lower bound** | **Upper bound** |
| 1 | PCTIM | Permanent impervious area | 0 | 5 |
| 2 | ADIMP | Saturated impervious area | 0 | 2 |
| 3 | UZTWM | Upper zone tension water maximum storage | 0.1 | 50 |
| 4 | LZTWM | Lower zone tension water maximum storage | 0.1 | 70 |
| 5 | LZFSM | Lower zone free water supplemental maximum storage | 0.1 | 100 |
| 6 | LZFPM | Lower zone free water primary maximum storage | 0.1 | 100 |
| 7 | LZSK | Lower zone supplemental withdrawal rate | 0.1 | 3.8 |
| 8 | SCF | Snowfall correction factor | 0.5 | 1.5 |
| 9 | REXP | Percolation equation exponent | 0.1 | 3.5 |
| 10 | UZK | Upper zone free water withdrawal rate | 0.1 | 3.5 |
| 11 | Q0CHN | Routing parameter | 0.5 | 4.5 |
| 12 | QMCHN | Routing parameter | 0.3 | 1.9 |

**5.5. Stepwise line search**

We use a stepwise line search approach to obtain the best fit parameter set. We use the square root of the mean square errors (i.e., the difference between observed and simulated flows) as an objective function. Stepwise line search follows the following steps: (1) start with the a priori estimates of the hydrologic model parameters; (2) with the rest of the parameter estimates fixed to the a priori, increase or decrease the value of the first parameter by one step to the direction of decreasing objective function value; (3) with the first parameter now set to the new (or old, if the objective function value did not decrease) value, decrease or increase the value of the second parameter by one step to the direction of decreasing objective function value; (4) repeat Step 3 until the objective function is minimized with respect to each of all remaining parameters; (5) repeat Steps 2 through 4 until no further reduction in the objective function is realized.

**5.6. Precalibration**

The precalibration approach applies a screening criterion to a large ensemble of hydrologic model runs and removes the model runs that do not align with the observed data. The remaining model runs and their corresponding parameter sets will approximate the desired posterior and posterior predictive distributions. We begin with an initial ensemble of 5,000 model runs with input parameters settings selected from a 12-dimensional Latin hypercube design. From the initial ensemble, we select a subset of 165 runs that fulfilled the following screening criteria - hydrological model output (streamflow) at each observation date must be greater than 25% of the observed streamflow and less than 175% of the observed streamflow. We extract the corresponding parameter values and outputs to construct Figures 3 - 7.

**5.7. Evaluation Metrics**

We employ both deterministic and probabilistic metrics to assess the quality of model calibration. Specifically, the following four metrics are considered: Klinge-Gupta Efficiency (KGE), Brier skill score (BSS), mean continuous ranked probability skill score (CRPSS), and Relative Operating Characteristics (ROC) curve.

**Kling-Gupta Efficiency (KGE)**

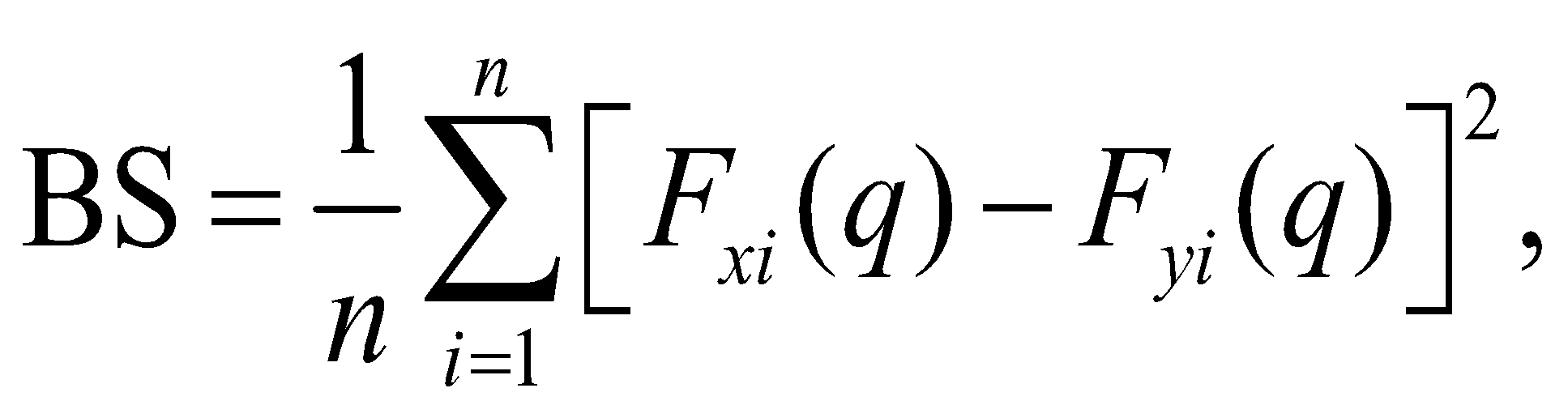
The KGE (Gupta et al., 2009) provides direct assessment of four aspects of streamflow time series, namely shape, timing, water balance and variability. It is widely used to indicate how well the model output fit the observations. It assesses the fit between modelled and observed values on a scale of –infinity to 1, with 1 indicating a perfect match and decreasing KGEs indicating decreasing fits between model results and observations. The KGE is computed as follows:

(1)

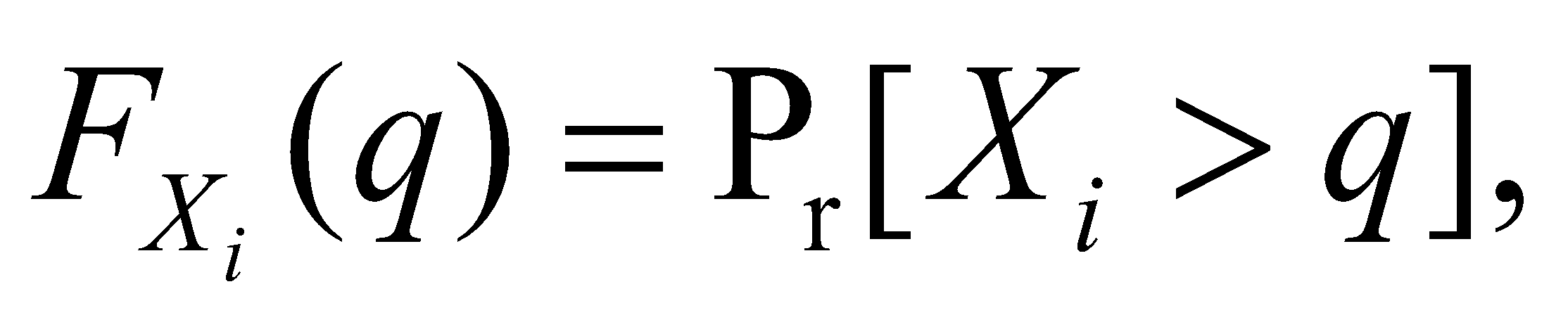
where is the linear correlation between observations and simulations, a measure of relative variability (ratio of simulated and observed standard deviation), and a bias term (ratio of simulated and observed means).

**Brier Skill Score (BSS)**

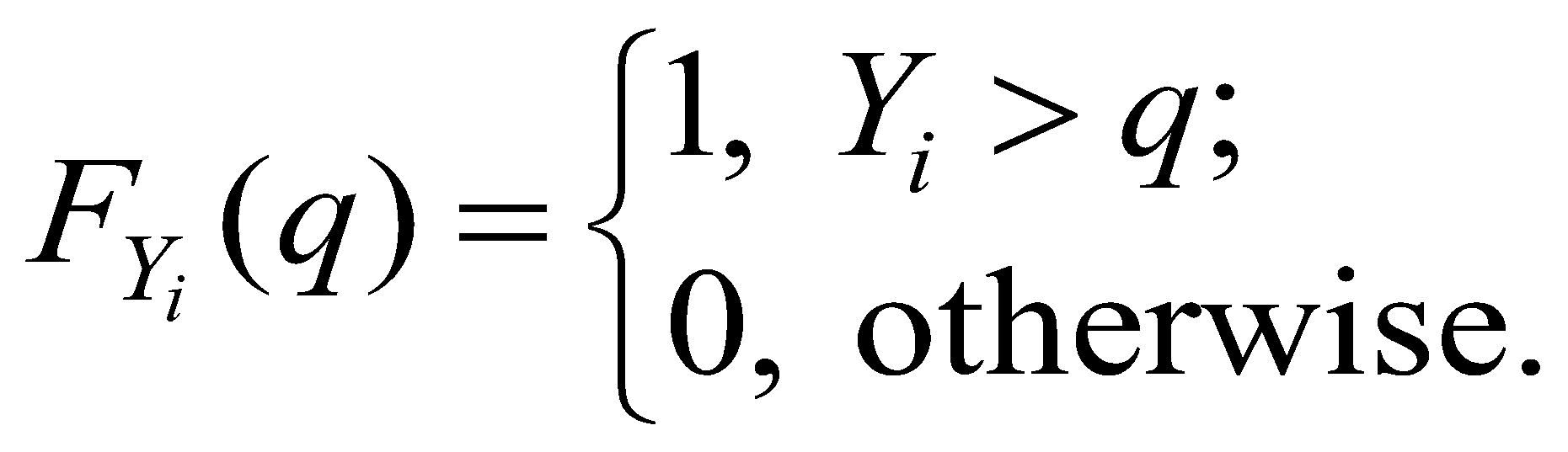
Brier score (BS) is analogous to the mean squared error, but where forecast is a probability, and the observation is either a 0 or 1 (Brown et al. 2010). The BS is given by

 (2)

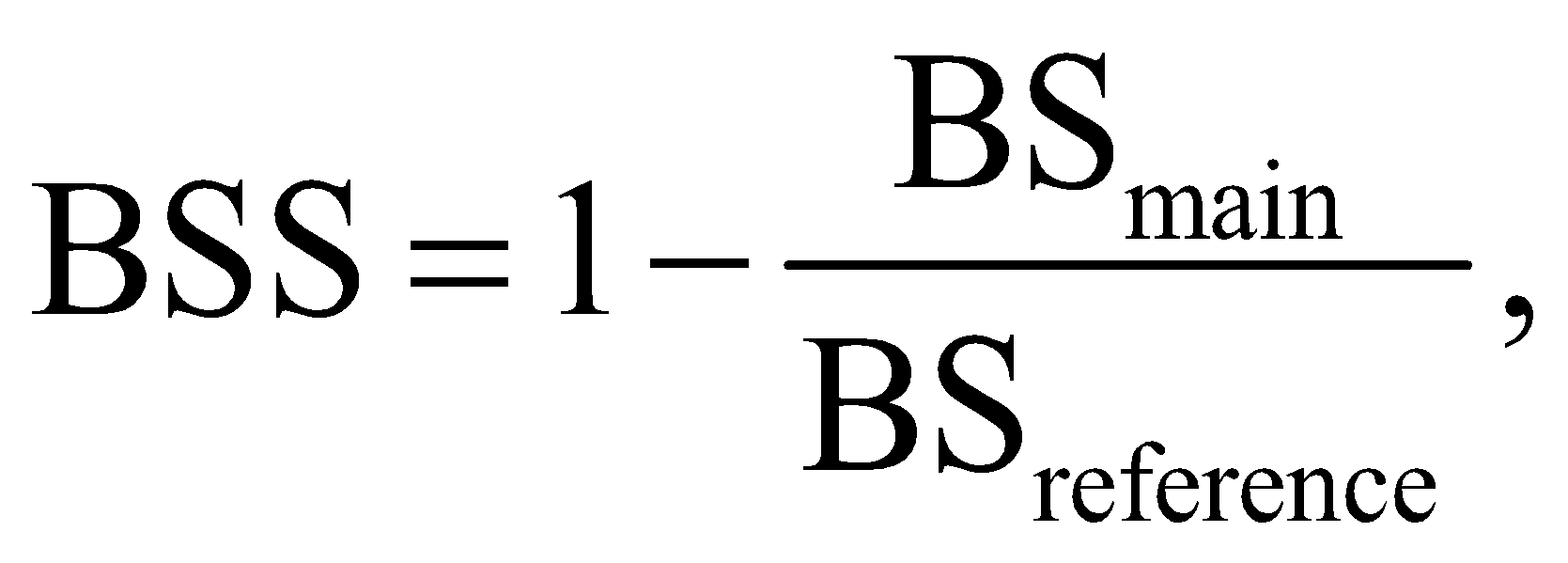
where the probability of *Xi* to exceed a fixed threshold (*q*) is

 (3)

*n* is again the total number of forecast-observation pairs, and

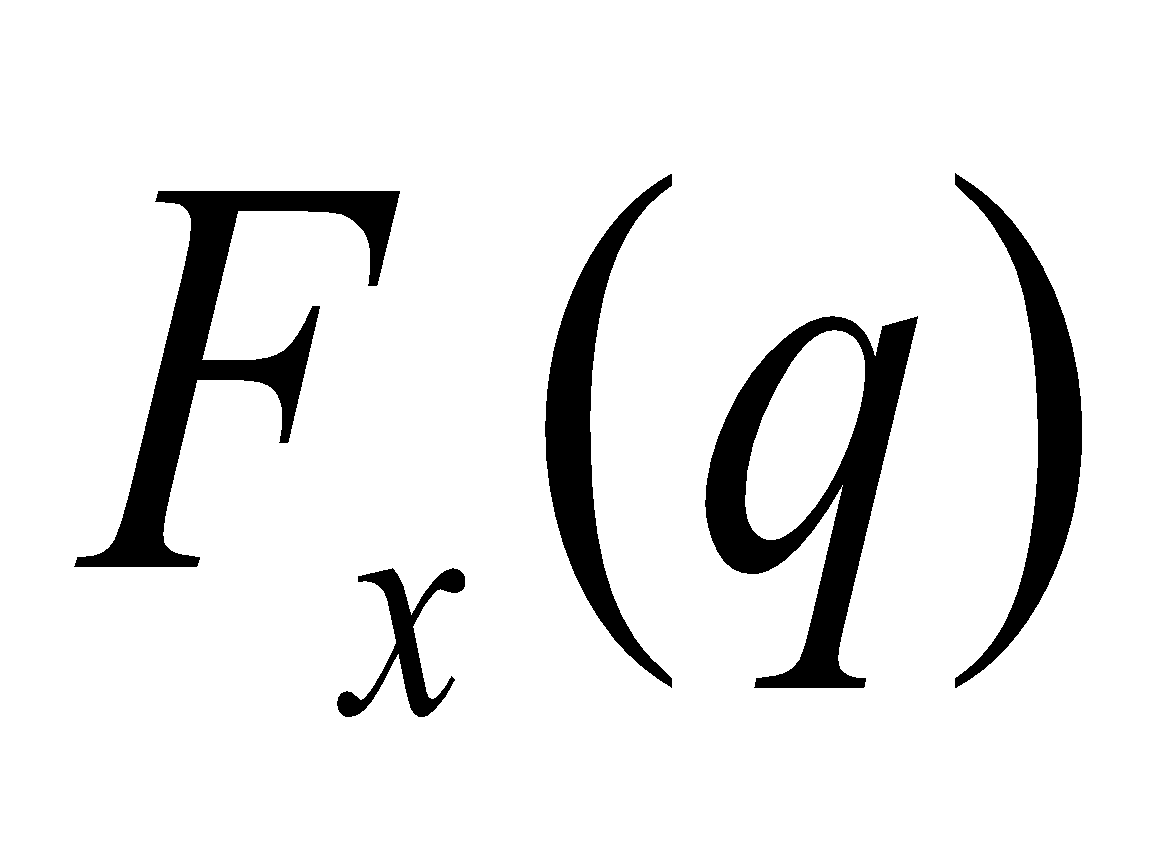
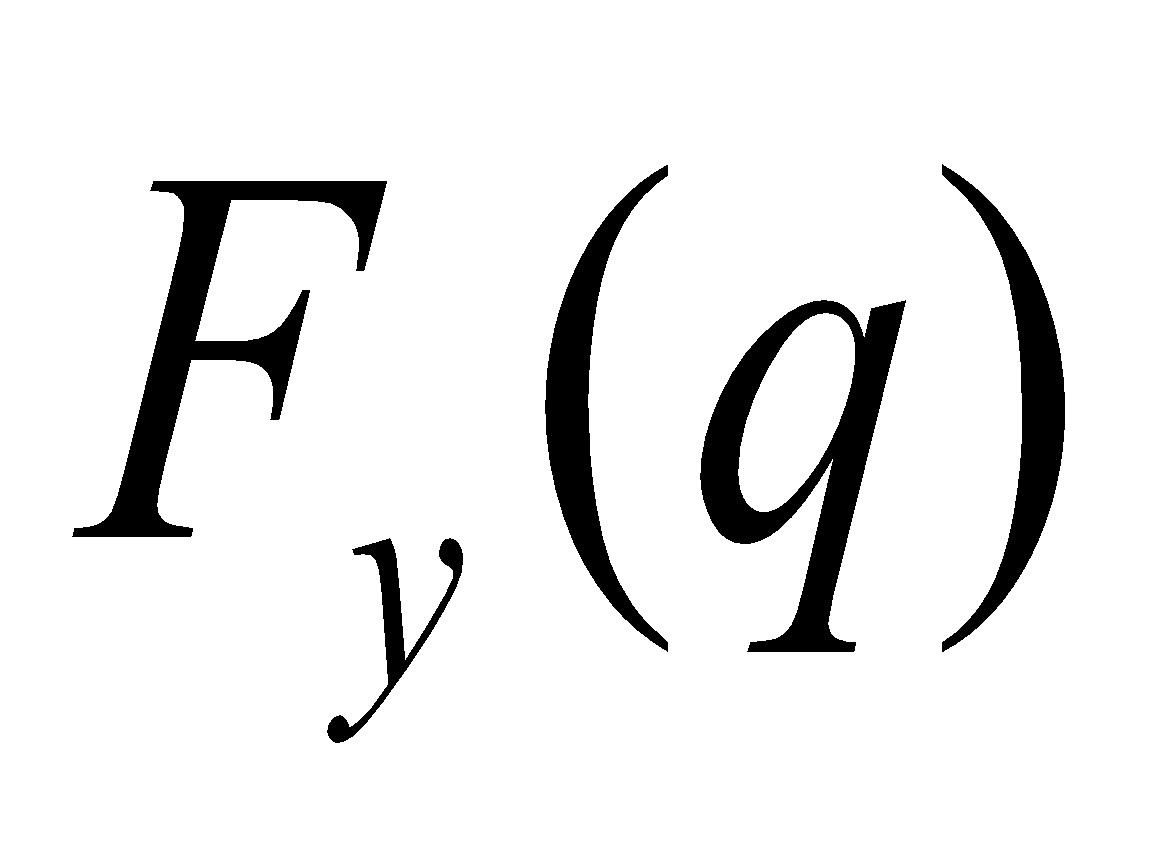
 (4)

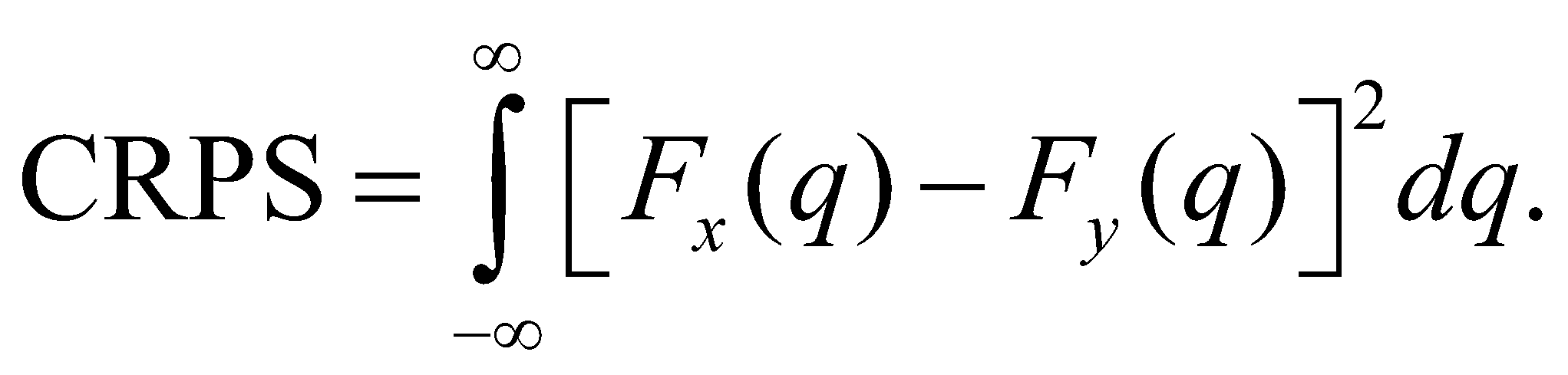
In order to compare the skill score of the main forecast system with respect to the reference forecast, it is convenient to define the Brier Skill Score (BSS):

 (5)

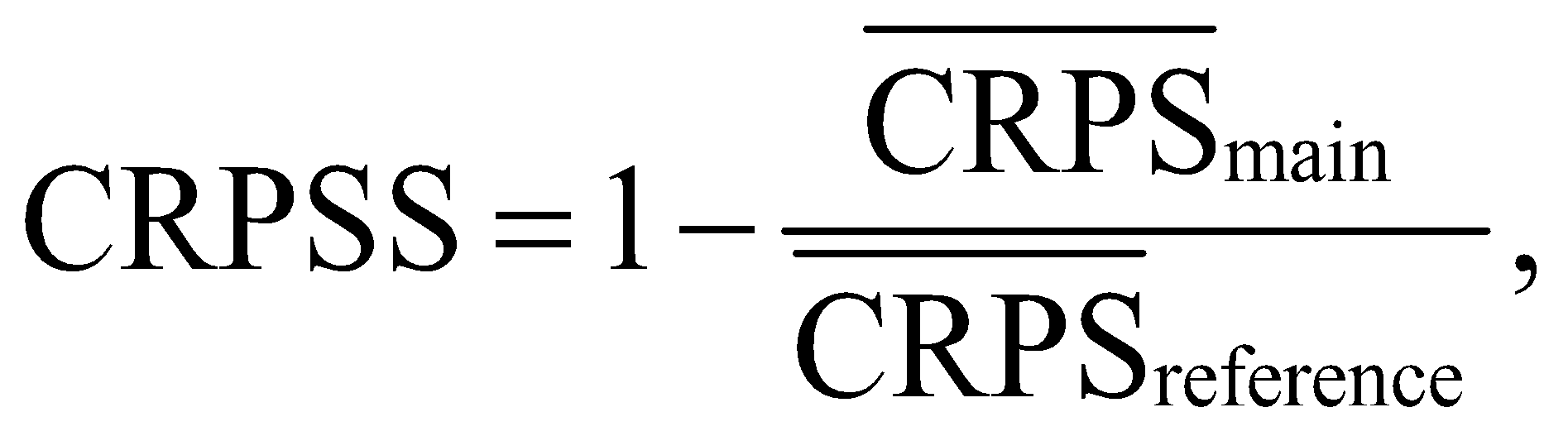
where BSmain and BSreference are the BS values for the main forecasting system (i.e., the system to be evaluated) and reference forecasting system, respectively. Any positive values of the BSS, from 0 to 1, indicate that the main forecasting system performed better than the reference forecasting system. Thus, a BSS of 0 indicates no skill and a BSS of 1 indicates perfect skill.

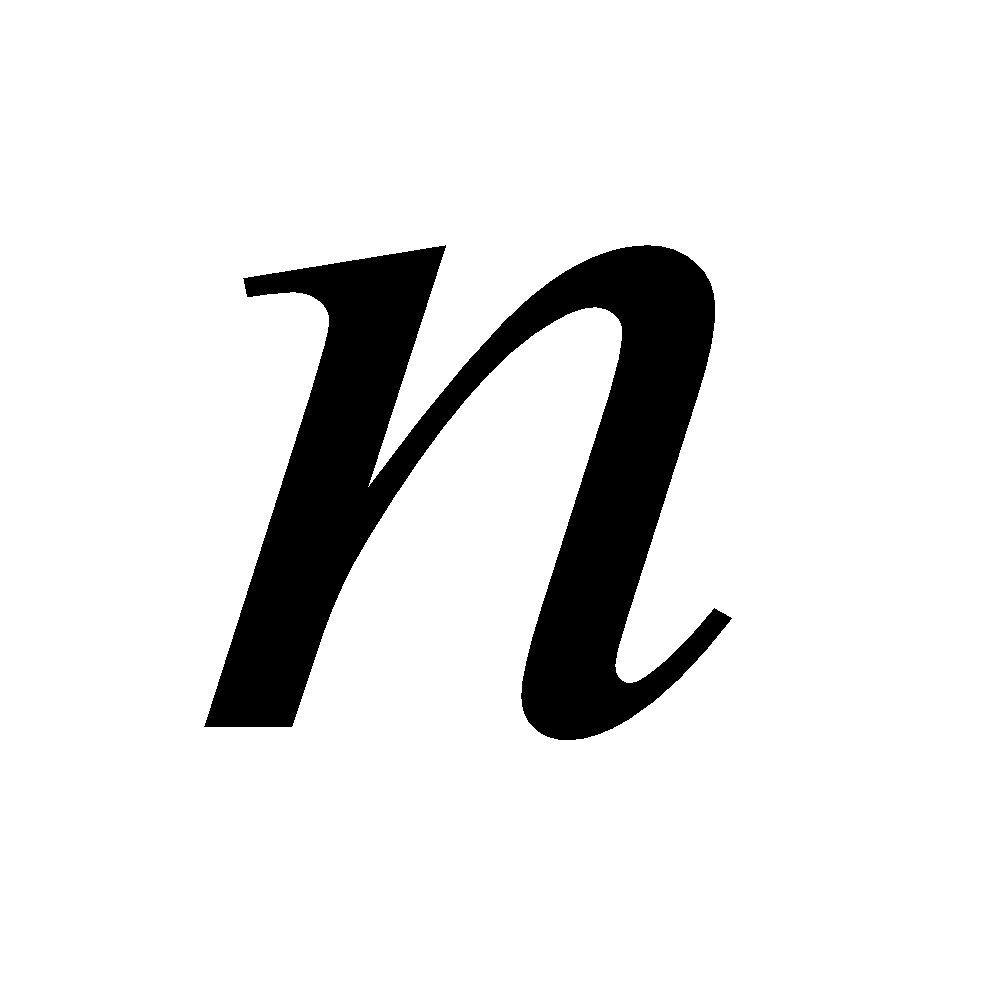
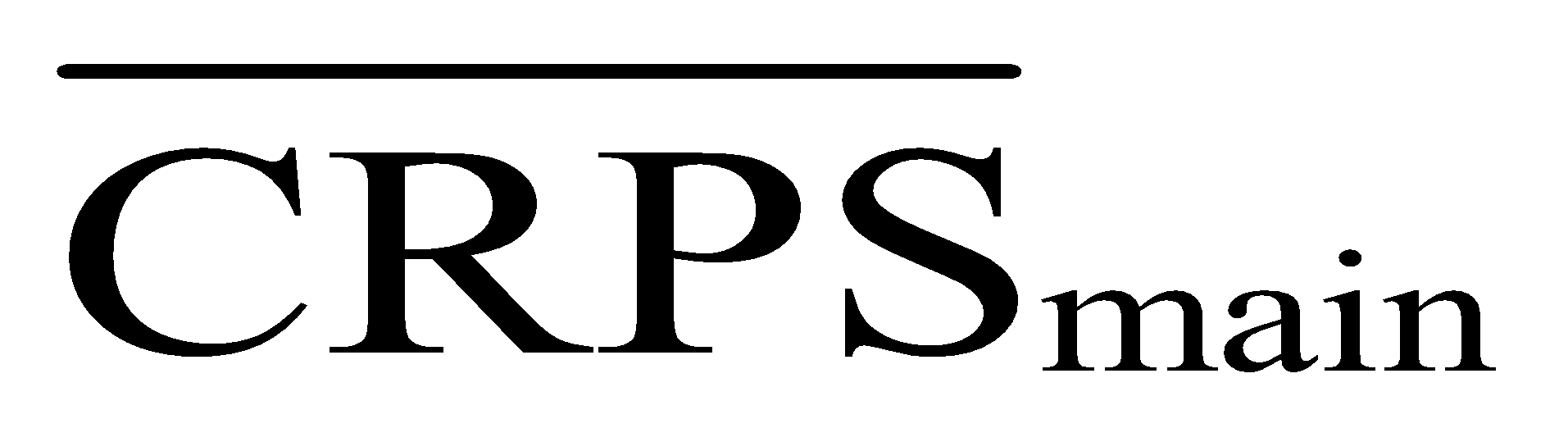
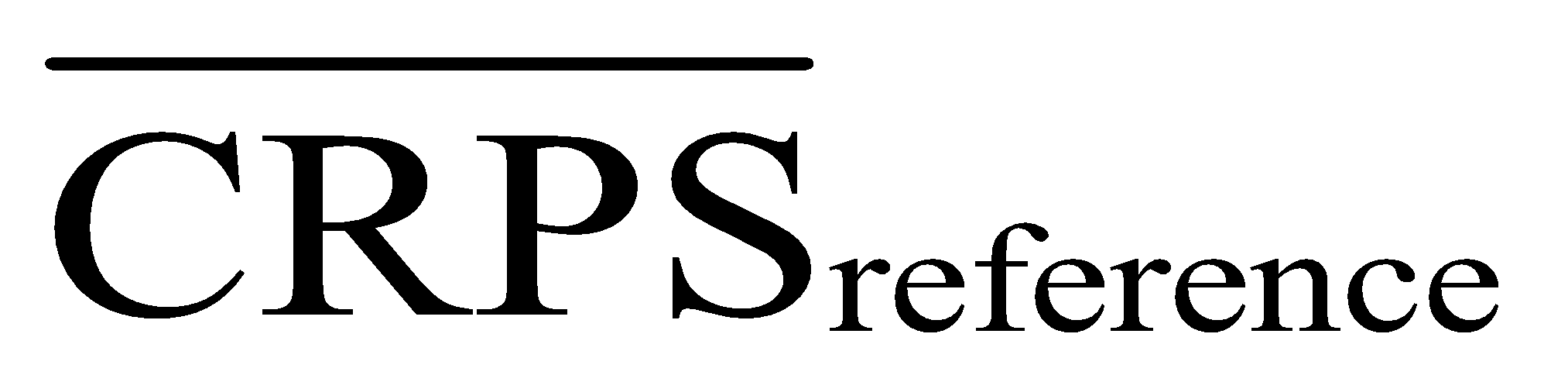
**Mean Continuous Ranked Probability Skill Score (CRPSS)**

The Continuous Ranked Probability Score (CRPS), which is less sensitive to sampling uncertainty, is used to measure the integrated square difference between the cumulative distribution function (cdf) of a forecast, , and the corresponding cdf of the observation, . The CRPS is given by

 (8)

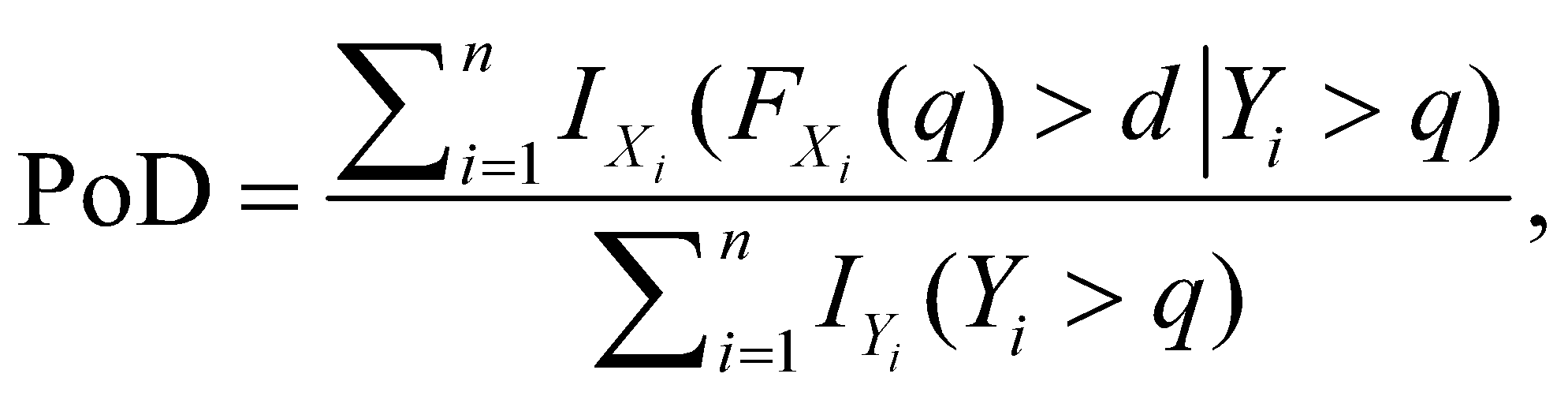
To evaluate the skill of the main forecasting system relative to the reference forecast system, the associated skill score, the Mean Continuous Ranked Probability Skill Score (CRPSS), is defined as:

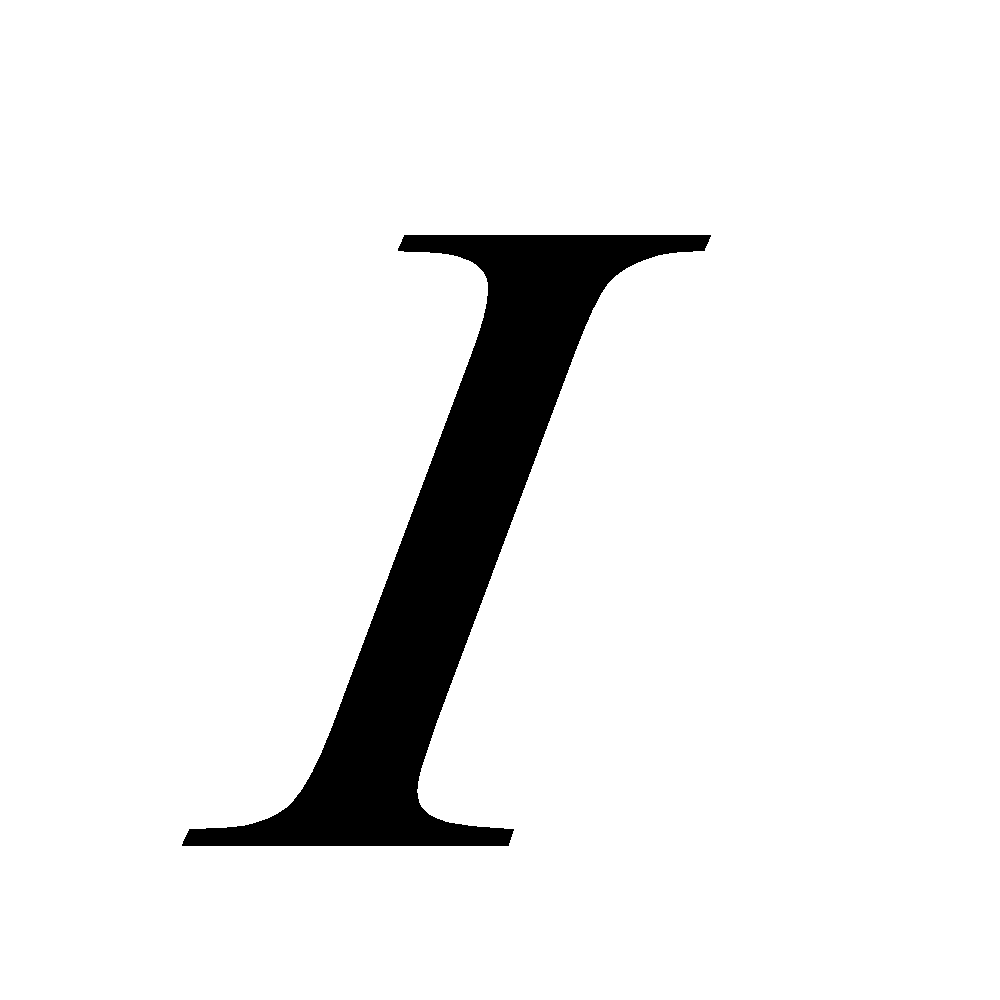
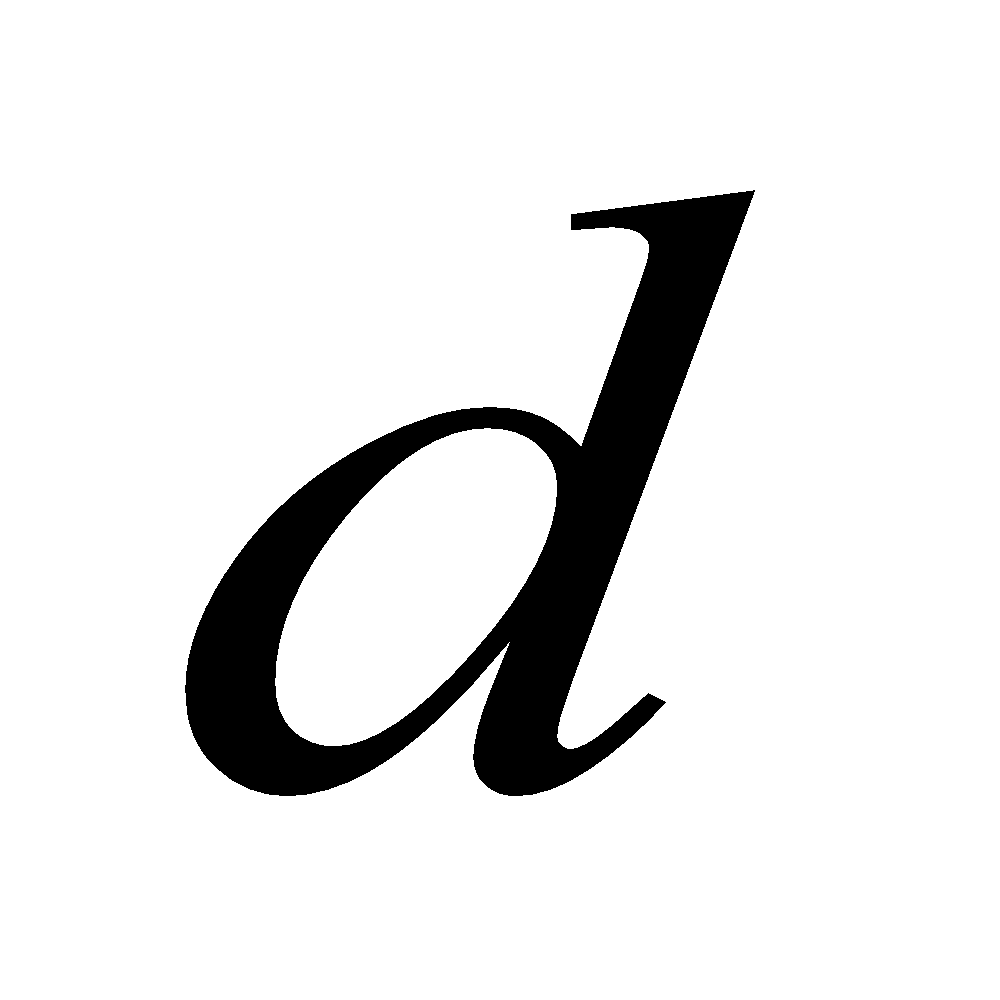
 (9)

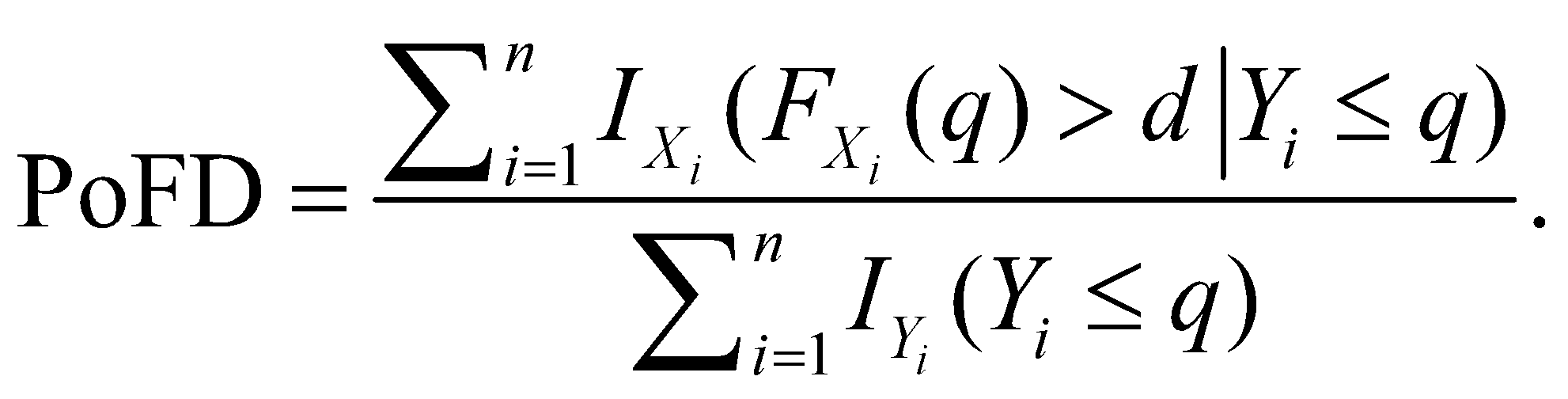
where CRPS is averaged across  pairs of forecasts and observations to calculate mean CRPS of the main forecast system () and reference forecast system (). The CRPSS ranges from -∞ to 1, with negative scores indicating that the system to be evaluated has worse CRPS than the reference forecasting system, while positive scores indicate a higher skill for the main forecasting system compared to the reference forecasting system, with 1 indicating perfect skill.

**Relative operating characteristic (ROC) curve**

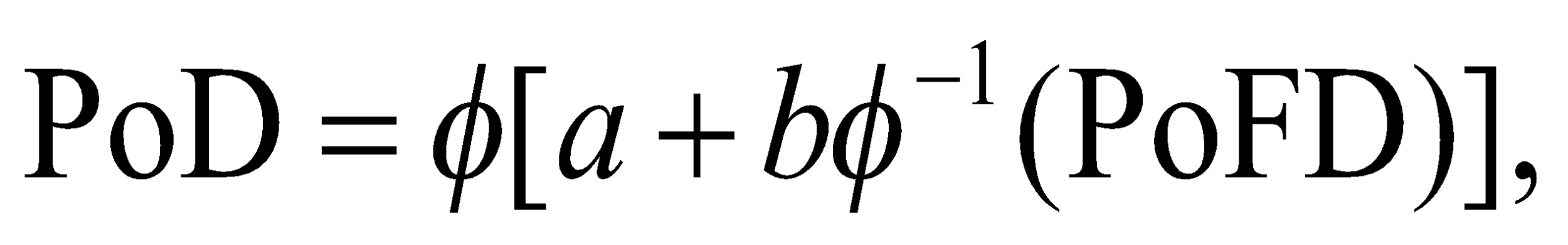
The ROC curve is a measure of the quality of probability forecasts that relates the probability of detection (PoD) or true alarm to the corresponding probability of false detection (PoFD) or false-alarm rate, as a decision threshold is varied across the full range of a continuous prediction quantity. For a particular threshold, the PoD is given by

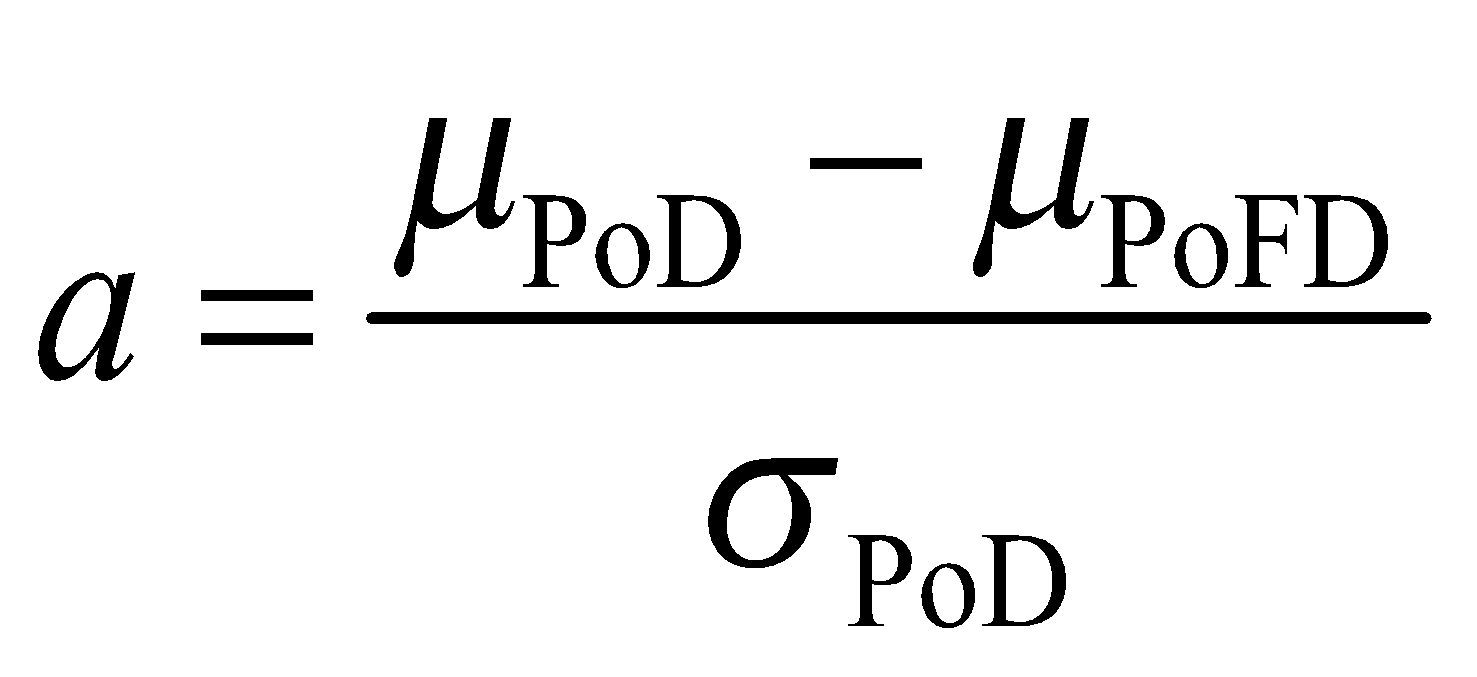
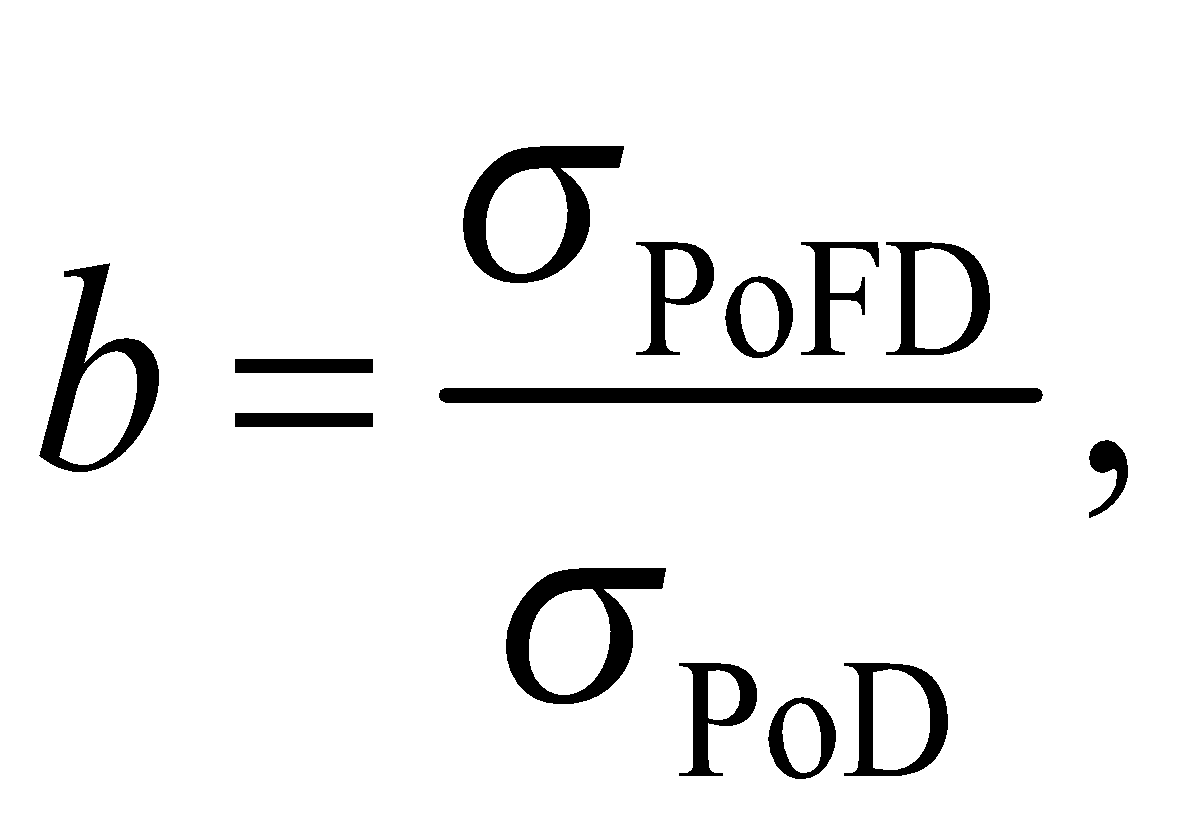
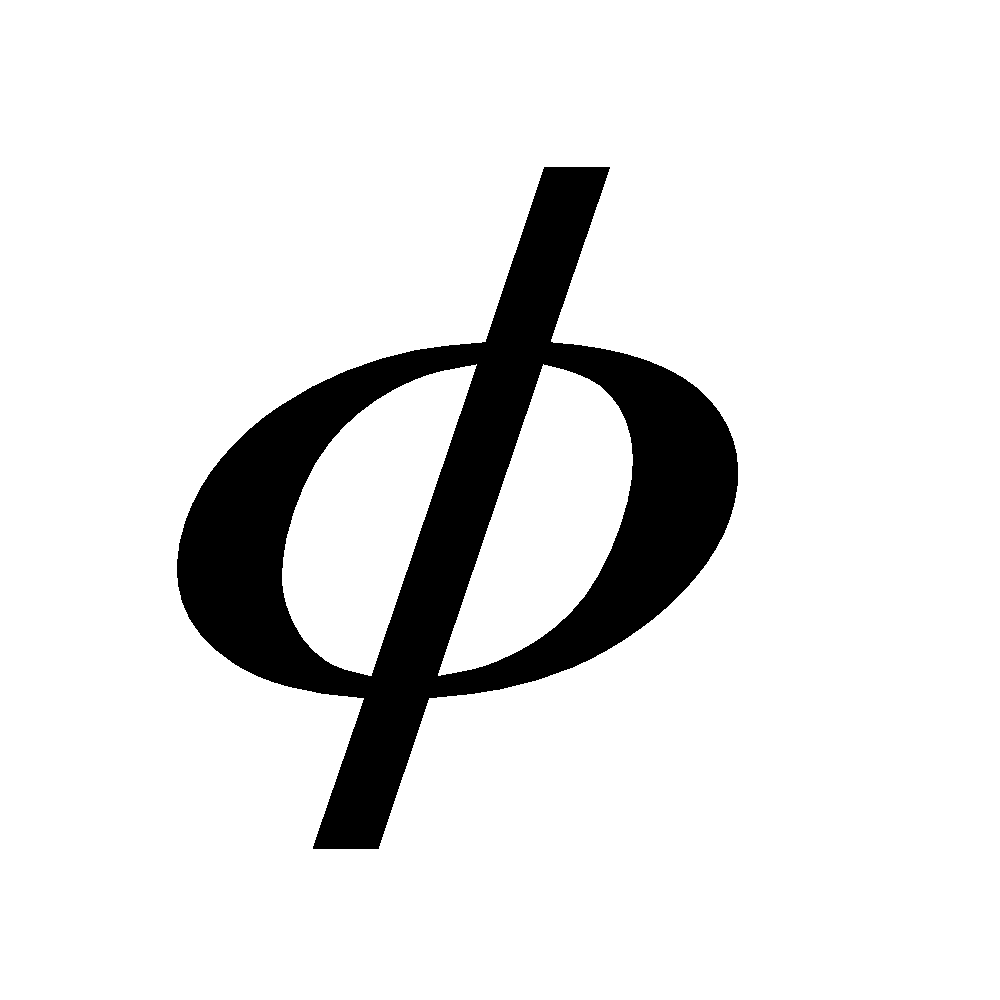
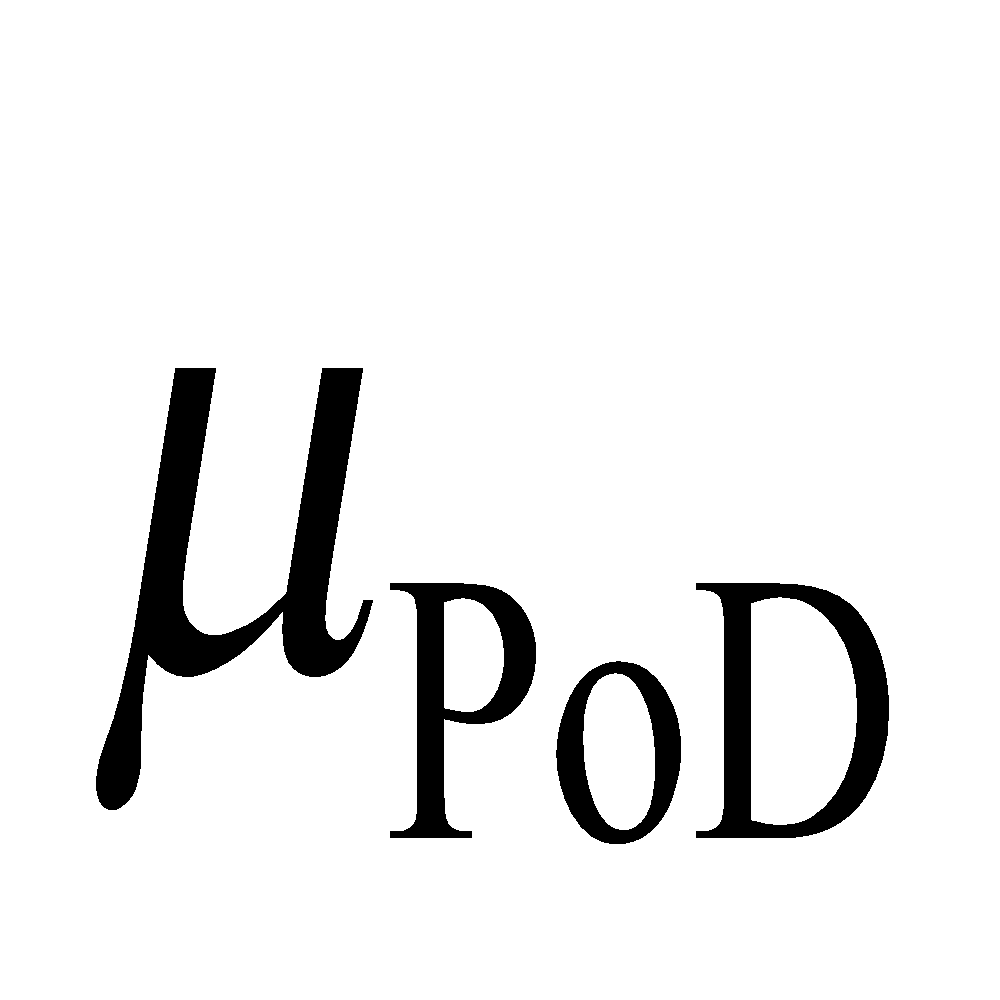
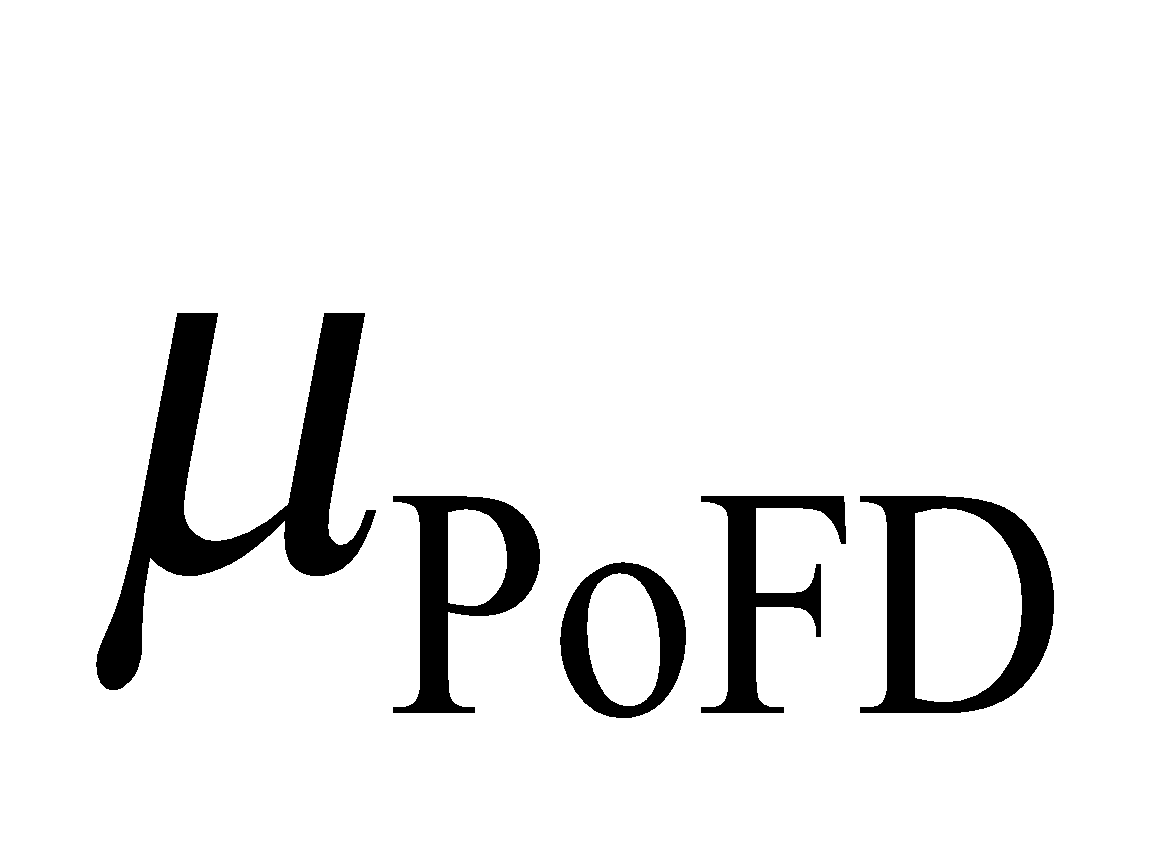
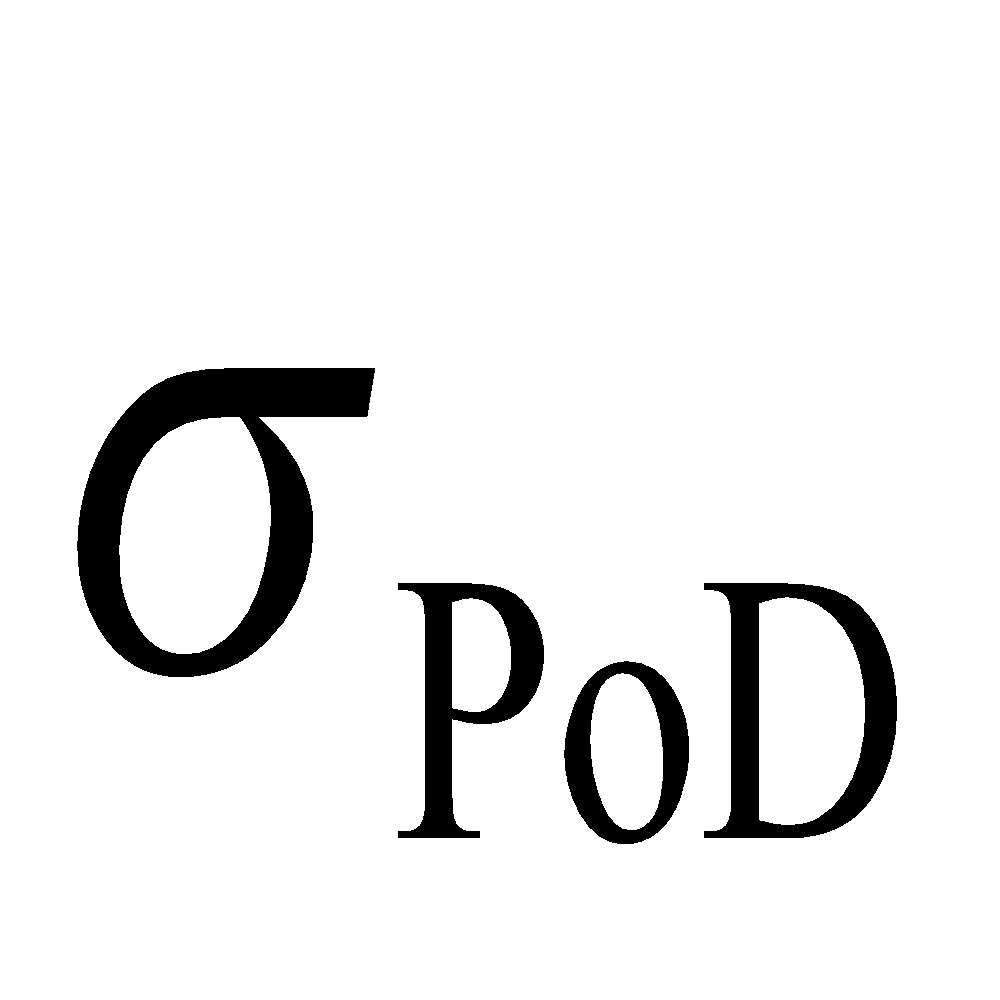
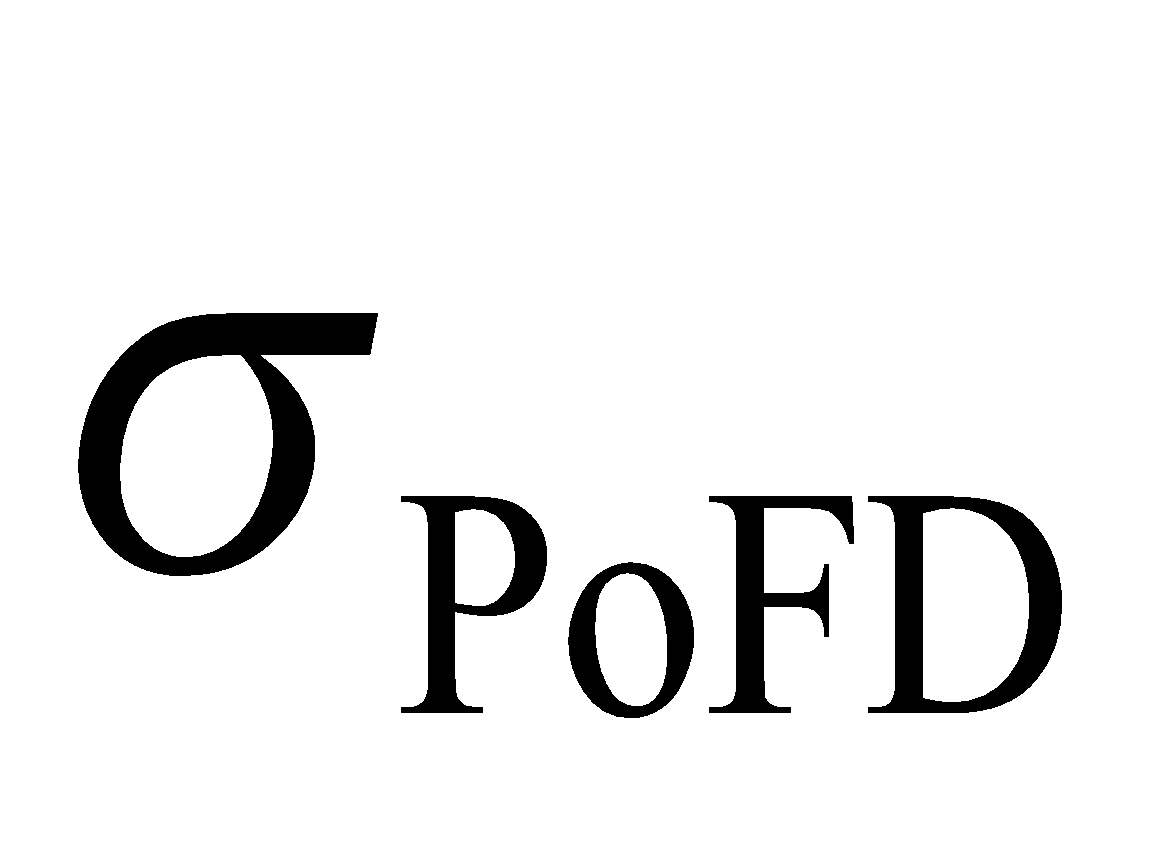
 (10)

where  denotes the indicator function and  denotes the probability threshold at which the event triggers some action. Similarly, the PoFD can be expressed as

 (11)

The relationship between PoD and PoFD is assumed bivariate normal such that

 (12)

where ,  and  is the cdf of the standard normal distribution.  and  are the means while  and  denote the standard deviations of the PoD and PoFD, respectively. The ROC curve plots the PoD (fraction of true alarms) against the PoFD (fraction of false alarms) for all possible values of the decision threshold, *d* [0,1], noting that an ensemble forecast is essentially a step function, with as many possible values of *d* as the number of ensemble members. The ROC score is based on the area underneath the ROC curve, which is normalized by the area under curve of the reference forecast.

**5.9. Flood Damage Estimates**

We calculate total damage to flood hazards. We consider 2,000 hypothetical houses to compute damage from flood hazards. Common vulnerability models are depth-damage functions that quantify the damages for a certain depth of water in a house. A common source of depth-damage functions in damage assessment studies in the U.S. is Hazard U.S. (HAZUS) provided by the Federal Emergency Management Agency (FEMA).

**Appendix A: FaMoS (FaMoS details goes into the Appendix)**

**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 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. constructed the particle-based calibration model. S.S. and B.L. led the calculations. I.S. performed a code review and edited the paper. 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 will be available through a publically accessible GitHub repository and under the GNU open-access license after acceptance in a peer-reviewed journal.

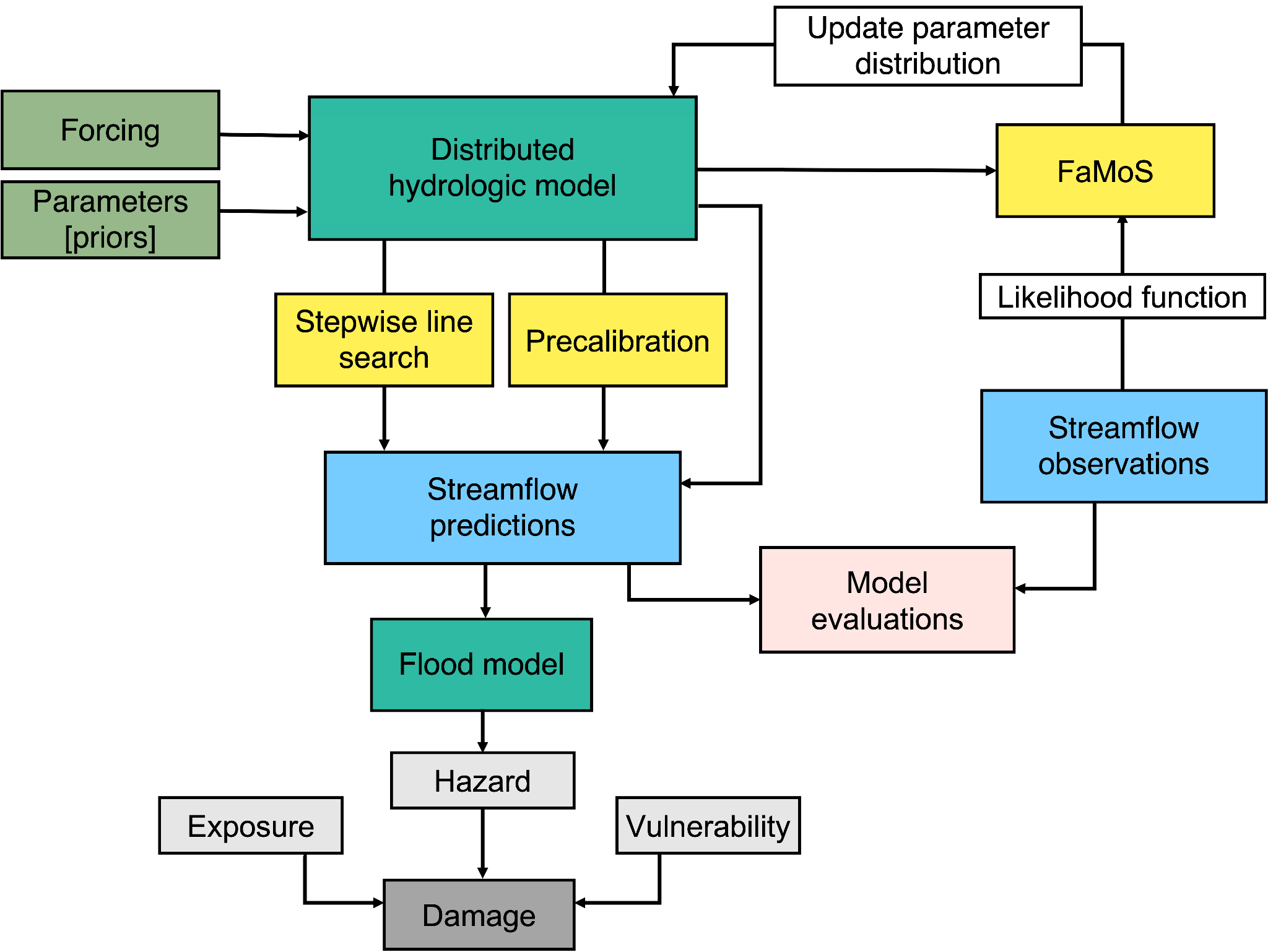
**Competing interests**

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

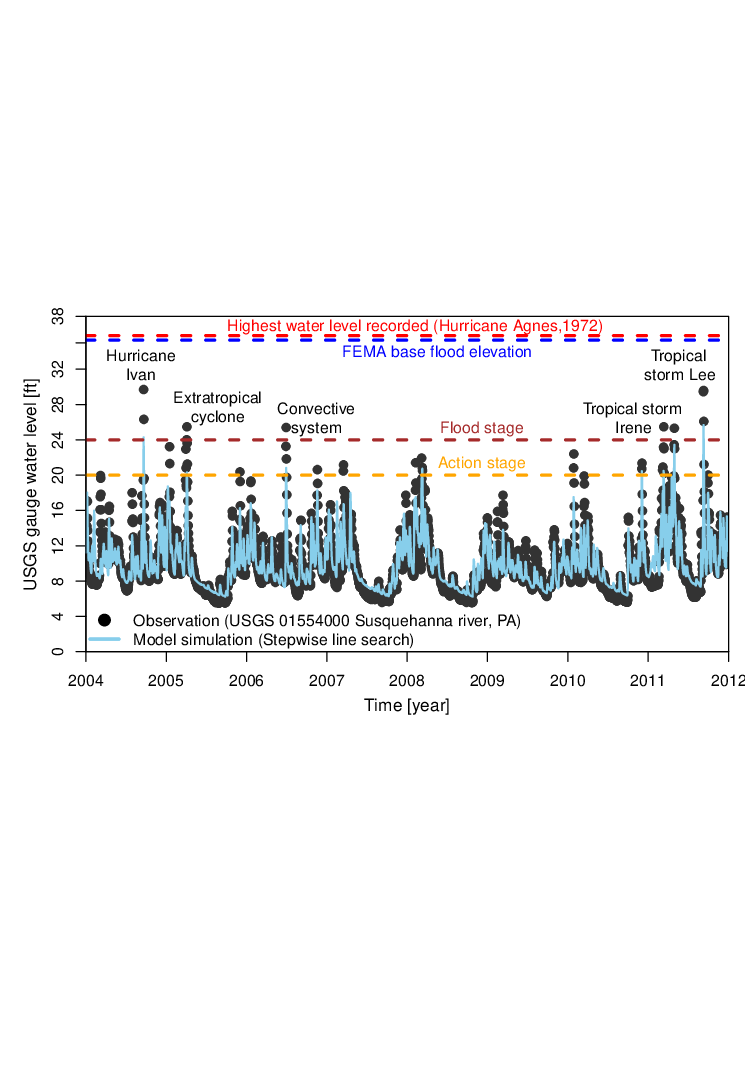
**Materials & Correspondence**

Correspondence and requests for materials should be addressed to the corresponding autho

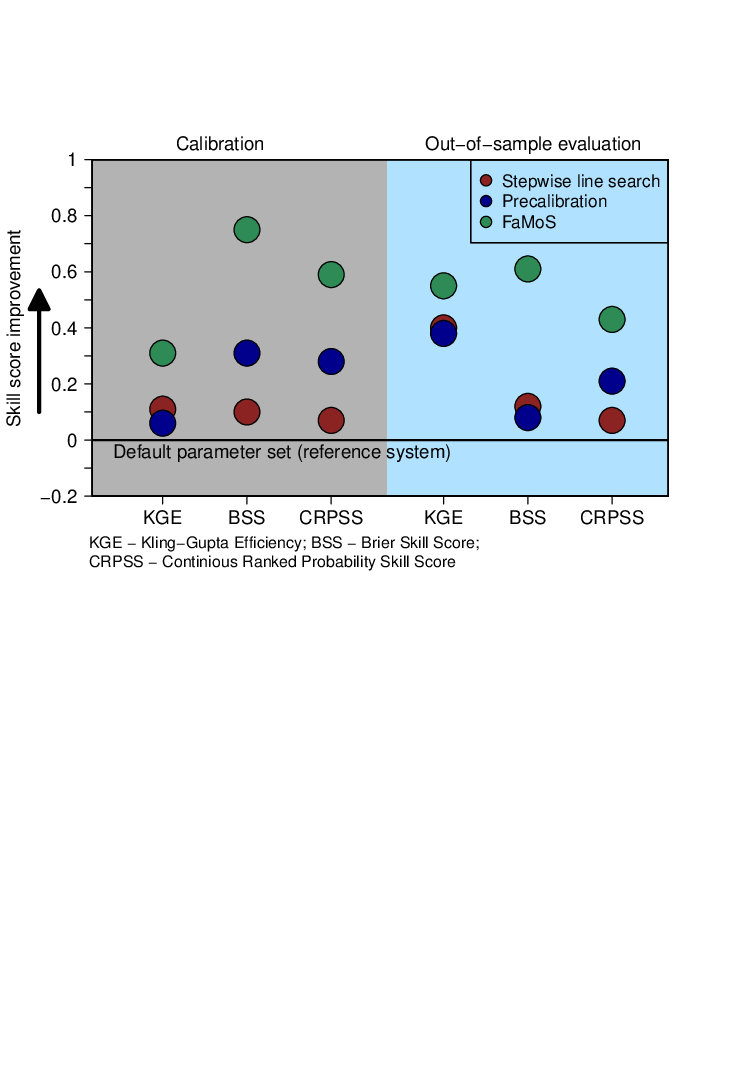
**List of Figures**



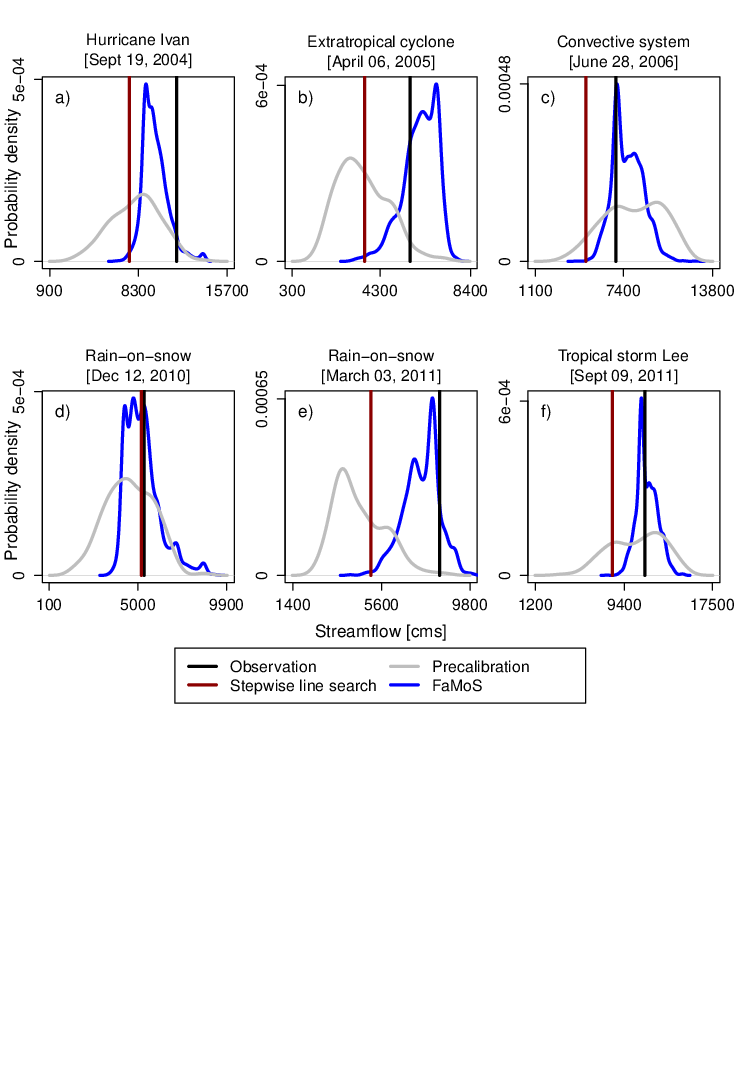
**Figure 1:** Diagrammatic representation of hydrological model calibration framework.



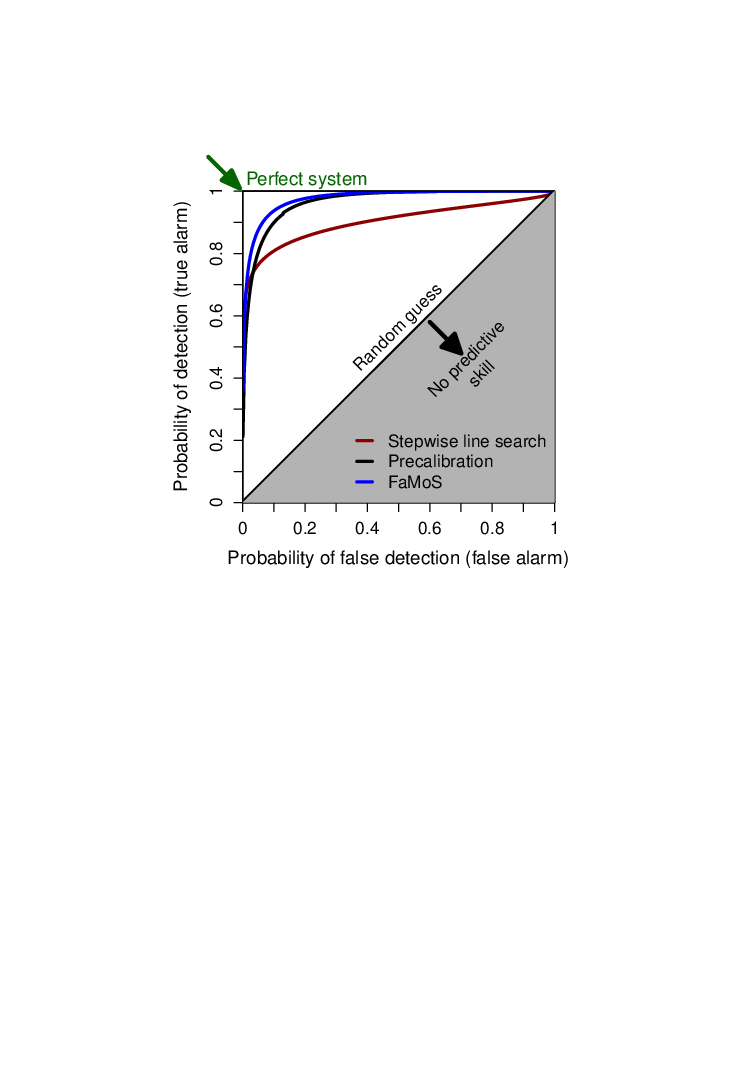
**Figure 2:** Historical time series of water level observation and model simulations obtained using sequential model calibration. We obtain the observation from the United States Geological Survey (USGS) gauge records for ID 01554000 located upstream of Selinsgrove, Pennsylvania, USA.



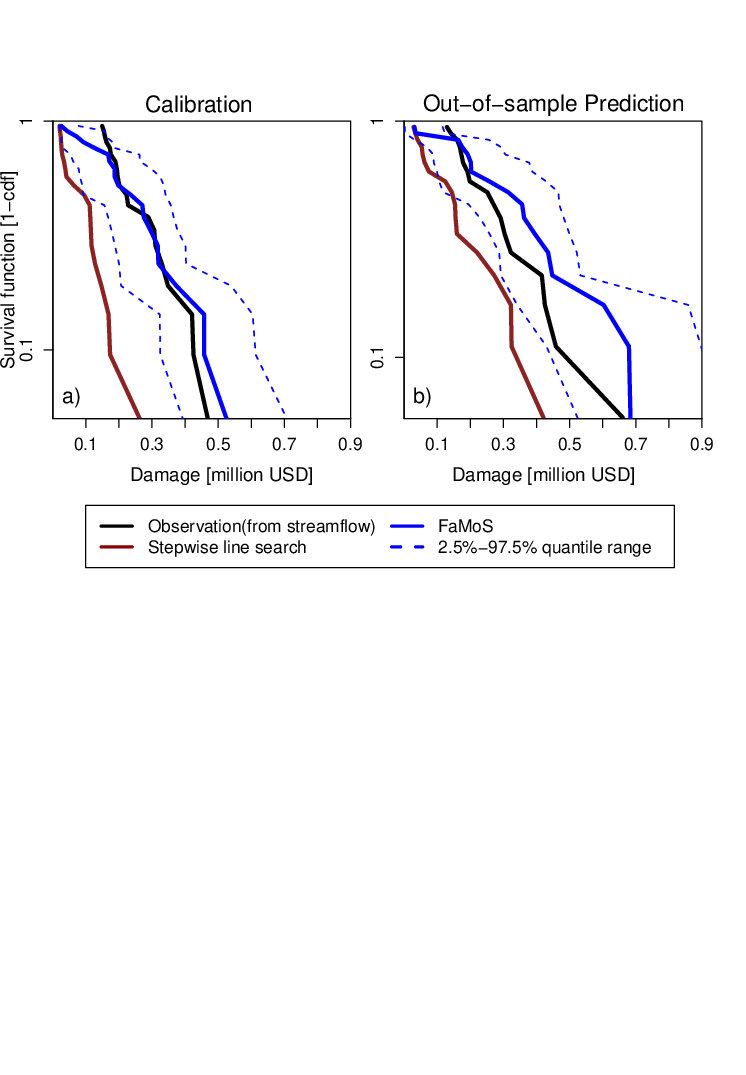
**Figure 3:** Performance metrics for hydrological model calibration and out-of-sample prediction. We compute Kling-Gupta Efficiency (KGE), and Brier skill score (BSS), and Continuous ranked probability skill score (CRPSS). All the metrics are computed with reference to the default parameter set available 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.



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

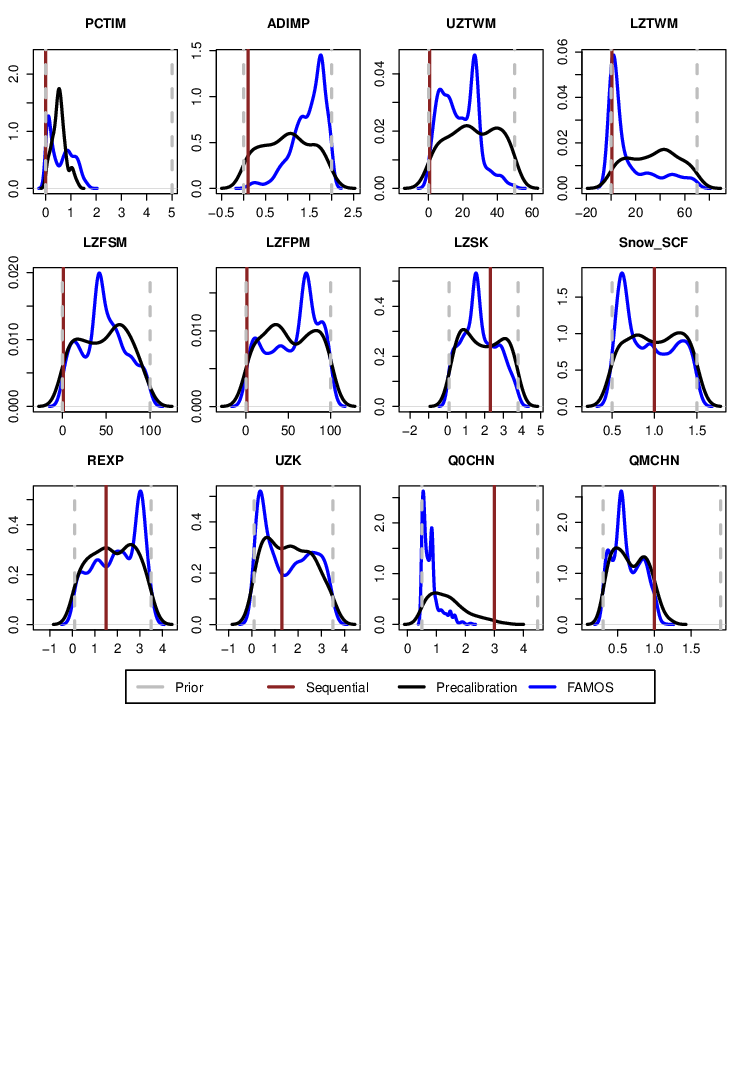


**Figure 5:** Relative operating characteristics (ROC) curve for different calibration approaches. ROC curve plots the probability of detection against the probability of false detection for a range of forecast probability levels. A larger area under the ROC curve represents a more skillful prediction, 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.



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

**Supplemental Figures:**



**Figure S1:** Prior and posterior distribution of hydrological model parameters.

**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/](http://paperpile.com/b/9lxFNc/fQXtd)[10.1002/2016ef000485](http://dx.doi.org/10.1002/2016ef000485)

[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/](http://paperpile.com/b/9lxFNc/JLmo8)[10.1029/2020wr028673](http://dx.doi.org/10.1029/2020wr028673)

[Chester, M. V., Shane Underwood, B., & Samaras, C. (2020). Keeping infrastructure reliable under climate uncertainty. *Nature Climate Change*. https://doi.org/](http://paperpile.com/b/9lxFNc/nUc7o)[10.1038/s41558-020-0741-0](http://dx.doi.org/10.1038/s41558-020-0741-0)

[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/](http://paperpile.com/b/9lxFNc/SiyT)[10.1029/2008wr006824](http://dx.doi.org/10.1029/2008wr006824)

[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.](http://paperpile.com/b/9lxFNc/u3Xxf)

[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.](http://paperpile.com/b/9lxFNc/4ELbE)

[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/](http://paperpile.com/b/9lxFNc/E8cn)[10.1002/2017wr020528](http://dx.doi.org/10.1002/2017wr020528)

[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.](http://paperpile.com/b/9lxFNc/9gFVx)

[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.](http://paperpile.com/b/9lxFNc/uUnA)

[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.](http://paperpile.com/b/9lxFNc/yPBP)

[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.](http://paperpile.com/b/9lxFNc/NT0K)

[Murphy, A. H. (1970). THE RANKED PROBABILITY SCORE AND THE PROBABILITY SCORE: A COMPARISON. *Monthly Weather Review*, *98*(12), 917–924.](http://paperpile.com/b/9lxFNc/xrOj)

[Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology and Climatology*, *12*(4), 595–600.](http://paperpile.com/b/9lxFNc/XLdw)

[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/](http://paperpile.com/b/9lxFNc/48nY)[10.1029/2011wr011123](http://dx.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.](http://paperpile.com/b/9lxFNc/oByr2)

[Razavi, S., & Tolson, B. A. (2013). An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, *49*(12), 8418–8431.](http://paperpile.com/b/9lxFNc/aRP7D)

[Salas, J. D., Obeysekera, J., & Vogel, R. M. (2018). Techniques for assessing water infrastructure for nonstationary extreme events: a review. *Hydrological Sciences Journal*, *63*(3), 325–352.](http://paperpile.com/b/9lxFNc/72IVp)

[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/](http://paperpile.com/b/9lxFNc/5VPxv)[10.1029/2019ef001391](http://dx.doi.org/10.1029/2019ef001391)

[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/](http://paperpile.com/b/9lxFNc/1bLwF)[10.1175/JHM-D-20-0238.1](http://dx.doi.org/10.1175/JHM-D-20-0238.1)

[Van Liew, M. W., Arnold, J. G., & Bosch, D. D. (2005). Problems and potential of autocalibrating a hydrologic model. *Transactions of the ASAE. American Society of Agricultural Engineers*, *48*(3), 1025–1040.](http://paperpile.com/b/9lxFNc/mIBQy)

[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.](http://paperpile.com/b/9lxFNc/8gr4l)

[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.](http://paperpile.com/b/9lxFNc/qyH3w)

[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.](http://paperpile.com/b/9lxFNc/GrqrC)

[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/](http://paperpile.com/b/9lxFNc/XijHS)[10.1002/2017ef000607](http://dx.doi.org/10.1002/2017ef000607)

[Zarekarizi, M., Srikrishnan, V., & Keller, K. (2020). Neglecting uncertainties biases house-elevation decisions to manage riverine flood risks. *Nature Communications*, *11*(1), 5361.](http://paperpile.com/b/9lxFNc/rVyIU)

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.

Higdon, D., Lee, H., & Holloman, C. (2002). Markov chain Monte Carlo Based Approaches for Inference in Computational Intensive Problems. In Bayesian Statistics 7, Proceedings of the Seventh Valencia Meeting.

Mak, S., Sung, C. L., Wang, X., Yeh, S. T., Chang, Y. H., Joseph, V. R., ... & Wu, C. J. (2018). An efficient surrogate model for emulation and physics extraction of large eddy simulations. Journal of the American Statistical Association, 113(524), 1443-1456.

Park, S., Hwang, J. P., Kim, E. & Kang, H. (2009) A new evolutionary particle filter for the prevention of sample impoverishment. IEEE Transactions on Signal Processing, 13 (4), 801–809.

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

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

Brunner, M. I., Melsen, L. A., Wood, A. W., Rakovec, O., Mizukami, N., Knoben, W. J. M., & Clark, M. P. (2020). Flood hazard and change impact assessments may profit from rethinking model calibration strategies. Hydrology and Earth System Sciences Discussions, 1–19.

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>

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.

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

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.

Kaipio, J. P., Kolehmainen, V., Somersalo, E., & Vauhkonen, M. (2000). Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse problems*, *16*(5), 1487.

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.

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. Annals of Applied Statistics. <https://doi.org/10.1214/19-aoas1305>

Li, T., Sattar, T. P., Han, Q., & Sun, S. (2013, July). Roughening methods to prevent sample impoverishment in the particle PHD filter. In Proceedings of the 16th international conference on information fusion (pp. 17-22). IEEE.

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.

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.

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.

Salas, J. D., Obeysekera, J., & Vogel, R. M. (2018). 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

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., 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

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.

Tang, Y., Reed, P., van Werkhoven, K., & Wagener, T. (2007). Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis. Water Resources Research, 43(6). https://doi.org/10.1029/2006wr005813

Thorstensen, A., Nguyen, P., Hsu, K., & Sorooshian, S. (2016). Using Densely Distributed Soil Moisture Observations for Calibration of a Hydrologic Model. Journal of Hydrometeorology, 17(2), 571–590.

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.

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

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

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.