Supplementary Material for

Breaking Down the Computational Barriers to Real-Time Urban Flood Forecasting

V. Y. Ivanov*, D. Xu*, M. C. Dwelle, K. Sargsyan, D. B. Wright, N. Katopodes, J. Kim, V. N. Tran, A. Warnock, S. Fatichi, P. Burlando, E. Caporali, P. Restrepo, B. F. Sanders, M. M. Chaney, A. M. B. Nunes, F. Nardi, E. R. Vivoni, E. Istanbulluoglu, G. Bisht, R. L. Bras

*Corresponding author email: ivanov@umich.edu, donghui.xu@pnnl.gov

This file includes:

Text S1. Data on floods, fatalities, and Houston flooding event.

- According to the flooding fatalities reports from NOAA, on average about 50% of the fatalities over the past 20 years were classified as 'in vehicle' and associated with vehicles trying to cross flooded roadways; 27% were classified as 'in water' where an individual was swept away by flood waters (see Table S1).
- Extreme floods (Figure S1c) are defined as events exceeding *Flood Magnitude* (FM) of 6, where FM is computed as the common logarithm of the product of flood duration, severity, and affected area [*Brakenridge*, 2016]. All fitted lines in Figure S1b-c are smoothed trends obtained through locally weighted polynomial regression (LOESS) [*Cleveland and Devlin*, 1988] using weighted linear least squares.
- Inundation data set [*FEMA*, 2018] for the August 2017 flooding in Houston, Texas, was developed by the Federal Emergency Management Agency (FEMA) and includes

maximum depth. This product is a post-event analysis of maximum inundation without reference to the actual time of occurrence. The depths were generated using interpolation of digital elevation model, high water marks, and modeling results.

- River stage data were collected by the U.S. Geological Survey at gage 08074540. The maximum discharge estimated with high confidence for the case study watershed (i.e., the streamflow confirmed with in-situ measurements by) is ~2.8 m^3/s (100 ft^3/s). Streamflows above this value are extrapolated from the USGS calibrated stage-discharge relationship. We estimated uncertainties associated with this extrapolation using gage rating data and the "ISO/WMO" uncertainty assessment method described in *Kiang et al.* [2018] (see Figure S2).

Text S2. Computational performance and scalability of flood models

There are numerous ways to assess computational performance of a numerical model. To enable a cross-comparison with models of similar as well as reduced-order of complexity, we introduce *Core Seconds per Simulated Day - Normalized* (CSSD-N), η:

$$\eta = \tau_{SD}/N, \qquad (1)$$

where τ_{SD} represents the total CPU seconds required for a 24-hour simulation period and N is the total number of computational cells used by the model. As the formulation suggests, η is a performance metric normalizing simulation time per unit cell, which attempts to eliminate the effect of dimensionality of the modeled problem. For example, in the current study, there are 271,215 grid cells in the case study domain, and ~52 hours are required to simulate a 24-hour

flood event with the flood-resolving high-fidelity model (FHFM) run in a serial mode (Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz). This yields a value of the CSSD-N for the FHFM: $\eta_{FHFM} = 52 \times 3600/271{,}215 = 0.69\ [sec]$, or, more generally, the typical order of η_{FHFM} is expected to be in the range $\mathcal{O}(10^{-2})$ - $\mathcal{O}(10^{0})$ [*Noh et al.*, 2019; *Sanders and Schubert*, 2019], even if the computational performance can be improved by using numerical solutions of higher efficiency (see Text S8 for a summary of details of the flood model implementation). Although inundation simulation is typically carried out at fairly small scales of a local watershed, as illustrated in the case study, this can translate to a large computational burden. This is because urban floods need to be simulated at high spatial resolution (several meters, consistent with "human action scale", as specified in the narrative) and correspondingly smaller time steps are required for numerical stability. We assess the magnitude of η_{FHFM} to be about 2 orders of magnitude higher than that estimated for global climate models of similar level of complexity (see below) that are rarely run in real-time.

General Circulation Models (GCMs) or Earth System Models (ESMs) are increasingly used in projections of the evolution of the various earth system components, including the fluid dynamics of the atmosphere and oceans. While GCMs/ESMs are not the ultimate examples of computational complexity, they can provide a useful benchmark for expectations for flooding computations because of the similarity of governing equations and scalability on parallelized platforms as compared to flooding / hydrodynamic models. Their typical computational meshes representing the Earth system are in the order ~$\mathcal{O}(10^{6}$-$10^{8})$ cells – another feature that makes their application comparable to that of flood models that can reach the same number of computational cells for a mid-sized or heavily urbanized watershed. One key difference nonetheless is that GCMs are predominantly run offline a limited number of times (typically less

than a few dozens of times) in large computational centers with dedicated allotment of CPU resources.

For example, to evaluate computational performance of GCMs in the Coupled Model Intercomparison Project 6 *Balaji et al.* [2017] used *Core Hours per Simulated Year* (CHSY) defined as the product of run time for one simulated year and the number of cores allocated to the simulation. A typical number of grid cells in GCM simulation may vary from $\mathcal{O}$ (10^6) to $\mathcal{O}(10^8)$, requiring from $\mathcal{O}(10^2)$ to $\mathcal{O}(10^5)$ core hours to finish a simulation of 1-year period. To provide a comparison with the simulation performance for an urban flooding problem, we convert CHSY metric to CSSD-N using the following transformation: $\eta_{GCM} = CHSY/365 * 3600/N_{GCM}$, where N_{GCM} denotes the number of computational cells used in a GCM. Values of CHSY and N_{GCM} reported in *Balaji et al.* [2017] lead to η_{GCM} varying from $\mathcal{O}$ (10^{-4}) to $\mathcal{O}$ (10^{-2}), which are 2 orders of magnitude lower than η_{FHFM} estimated above.

It should be noted that GCM simulations are typically run offline or in some instances in weather forecasting applications of several day lead time [*Alfieri et al.*, 2013]. This is usually done using parallel methods at high-performance computing (HPC) centers using $\mathcal{O}(10^2)$ - $\mathcal{O}(10^3)$ cores. While the computational performance of a high-fidelity flood model is somewhat inferior to that of a typical GCM, it is unlikely to expect that the problem can be resolved using a similar 'brute-force' approach, i.e., by allocating many CPUs for flood simulation in real-time. Model parallelization can shorten the execution time, but there is evidence that simulation will still not be fast enough for real-time forecasting as the speedup gain is limited (see Figure S3; the scalability of the high-fidelity model used in this study is similar to others reported in *Artichowicz and Gasiorowski* [2019] and *Neal et al.* [2010]). From a practical point of view,

modeling every urban center subjected to flooding during an extreme event using a large number of cores would create a deficit of computational resources. Furthermore, no rigorous uncertainty quantification would be possible since this requires many more additional model simulations (dozens to thousands, depending on the number of considered uncertain inputs).

Text S3. Surrogate modeling

A general objective for a *surrogate* or *reduced-order model* (or *emulator*, *response-surface model*) is to simulate the behavior of a more computationally complex model. The reason for using a surrogate model is that it is computationally inexpensive compared to the original, complex model. The surrogate model can therefore be rigorously sampled for uncertainty propagation, parameter inference, or sensitivity analysis. However, this comes at a cost. If many uncertain inputs are taken, or there are high-order interactions between uncertain inputs in the computationally expensive model $\mathcal{M}$, then a significant effort will need to be expended to train the surrogate. Nonetheless, recent advancements have made constructing surrogates computationally efficient, enabling surrogate representation for models of increasing complexity.

There are multiple frameworks that fall into the class of surrogate models, e.g., Gaussian process (GP) models [*Kennedy and O'Hagan*, 2000; *Rasmussen and Williams*, 2006], artificial neural networks [*Ripley*, 1996], support vector machines [*Abe*, 2010], and polynomial chaos expansions (PCEs) [*Le Maître and Knio*, 2010; *Xiu and Karniadakis*, 2002]. These different classes of surrogates were often developed in parallel fields over the last few decades, with GP models and PCEs mainly being used in statistics and engineering disciplines such as computational fluid

dynamics, whereas neural networks and support vector machines were preferred in data-intensive applications.

This study uses PCEs to construct surrogate models. This methodology is preferred because the construction of PCEs offers multiple benefits: it retains interpretability of uncertain inputs, follows the physical constraints of the underlying physical model, and enables global sensitivity analysis for quantities of interest. The methodology was developed by *Wiener* [1938], but remained largely untouched until recently as advances in computational power made the method feasible for addressing real-world engineering challenges [*Marzouk et al.*, 2007; *Najm*, 2009; *Xiu and Karniadakis*, 2003; *Xiu and Tartakovsky*, 2004]. A limited comparison of the performance of PCE-based surrogates with those based on neural networks for a flooding problem is provided in *Xu* [2020].

PCEs are briefly summarized here. Consider a deterministic mathematical model $\mathcal{M}$ with uncertain inputs $\boldsymbol{X}$ that is used to predict some output quantity of interest $y = \mathcal{M}(\boldsymbol{X})$. If the inputs $\boldsymbol{X}$ are considered to be uncertain, then it is of interest to determine the probability distribution of the outputs y. If the deterministic model $\mathcal{M}$ is computationally expensive, the naïve random sampling of $\boldsymbol{X}$ and estimating distribution or any statistics of y is too computationally prohibitive due to the large number of model evaluations required. For efficient uncertainty representation and propagation with a wide class of random variables, one employs PCEs. Both inputs and outputs are written as polynomials of standard random variables. Typically, prior to seeing any observational data, one constrains the input parameters by expert-informed ranges. Therefore, inputs $\boldsymbol{X}$ are written as linear transformation of a standard uniform random variable U[-1,1]. In the following, we keep this linear transformation in mind, and

assume $\boldsymbol{X}$ is a d-dimensional vector of standard uniform random variables, without losing generality.

The model's outputs can then be approximated as an expansion with respect to orthogonal Legendre bases $\Psi_k(\boldsymbol{X})$ [*Xiu and Karniadakis*, 2002]:

$$y \approx \widetilde{\mathcal{M}}(\boldsymbol{X}) = \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{X}). \qquad (2)$$

The goal is to obtain the right-hand side of the above equation, with c_k being the deterministic coefficients. Once these coefficients are known, one can calculate the distribution of the model output y as induced from the uncertain inputs $\boldsymbol{X}$. The coefficients can be solved for using Gaussian quadrature, regression, or Bayesian approaches [*Sargsyan et al.*, 2014]. Once known, the right side of the model can be evaluated against uncertain inputs $\boldsymbol{X}$ to get the distribution of outputs y. Additionally, due to the orthogonality of $\Psi_k(\boldsymbol{X})$, one can estimate sensitivity indices with respect to input parameters (components of $\boldsymbol{X}$), analytically.

In general, surrogate model ($\widetilde{\mathcal{M}}(\boldsymbol{X})$) construction may be challenged by overfitting in cases when the model capacity is much larger than the number of available high-fidelity simulations. In the PCE context, this is manifested in situations when there is a large number of uncertain inputs and not enough training simulations due to the computational burden of the high-fidelity model. However, the simple parametric form of PCE makes it less prone to overfitting than other surrogate model methodologies. Furthermore, sparse learning approaches, such as Bayesian compressive sensing [*Dwelle et al.*, 2019; *Sargsyan et al.*, 2014], facilitate adaptive selection of only relevant polynomial terms in the PCE, effectively enforcing the Occam's razor principle and further reducing the likelihood of overfitting.

Errors of surrogate models are generally tracked with a hold-out set of high-fidelity model simulations. The associated variance is compared with variance contributions produced by parametric sensitivity analysis, leading to improved decisions to help direct the efforts in terms of improving predictive uncertainties. Such improvements range from focusing on certain model parameters, regions, and/or times to improving surrogate forms or having a larger number of high-fidelity training simulations.

Text S4. Inference

This section describes an important structural element of the proposed framework, which for the sake of simplicity has been omitted from the case study. However, it has been previously shown to be effective in merging uncertain data and models in the context of Bayesian inference aided by surrogate modeling [e.g., *Sargsyan et al*., 2015; *Dwelle et al*., 2019; *Sargsyan et al.*, 2019]. Specifically, 'inverse modeling' is a class of problems for which the goal is to determine the "true" values of model inputs or parameters given observational data and a model [*Jaynes*, 2003]. In addition to parameter inference (i.e., in which case 'parameter' is understood in a typical geophysical context – an uncertain property of the media considered, or an uncertain constant in a closure scheme, such as hydraulic conductivity, resistance coefficient, etc.), one may perform inference on other model inputs that describe boundary or initial conditions of the system such as configuration of channel bathymetry or average initial soil moisture conditions. That is, 'inference' can encapsulate a broader range of possible scenarios than parameter estimation. The term "true" above is in quotes because while model inputs and parameters can carry physical meaning, they are subject to observational inaccuracies, scale of their

measurement and relevant process representation in the model, and overall model assumptions that affect their estimation. They thus represent "effective" actual values for a studied domain.

Conceptually, inference carried out with a high-fidelity model (aided with trained surrogates as described below) has the additional benefit of yielding results informative for studies and analyses with other models or another version of the same model. Consider two formulations (or versions) of high-fidelity models, A and B. If one performs an inverse modeling for a parameter representing surface roughness coefficient using model A, the posterior value of this coefficient will not necessarily fit well the observed data when used in model B. Nonetheless, obtaining the value for model A would be valuable as it provides more information about the parameter/input and therefore the watershed of interest, while also adjusting model A for future studies.

The inference part of the uncertainty quantification workflow is shown in Figure S4. Given a suite of results from a model $\mathcal{M}$ and its constructed polynomial surrogate $\widetilde{\mathcal{M}}$, one can infer which values of uncertain inputs $\boldsymbol{X}$ are *most likely* to provide results that match an observed quantity. The advantage of this approach is that $\widetilde{\mathcal{M}}$ enables very efficient inverse analysis [*Marzouk and Xiu*, 2009]. Surrogate models with dimension reduction as presented here provide an approach enabling faster computation, inversion, and the ability to solve the inverse problem on a larger set of uncertain model parameters and inputs.

Specifically, consider $\tilde{\boldsymbol{y}}$ to be a vector of observed data and $\boldsymbol{X}$ be the vector of uncertain model inputs. We also assume that the model gives an adequate approximation of the observed quantity. One can use Bayes' rule to compute the posterior distribution of input values conditional on the observed data [*Tarantola*, 2005]:

$$p(\boldsymbol{X} \mid \tilde{\boldsymbol{y}}) \propto \mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{X})\, p(\boldsymbol{X}),$$

where $p(\boldsymbol{X})$ is the prior distribution, $\mathcal{L}_{\mathcal{D}}$ is the likelihood function which represents the probability of obtaining the data given the set of inputs, essentially measuring the goodness-of-fit between model and data, while $p(\boldsymbol{X} \mid \tilde{\boldsymbol{y}})$ is the posterior distribution for $\boldsymbol{X}$, representing the updated probability density of $\boldsymbol{X}$, after taking the observed data into account.

To formulate a likelihood function, one must represent the discrepancy between the model and observations. Assuming that the difference between these are independent and identically distributed random variables with some marginal density p_{η}, the likelihood function can be written as:

$$\mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{X}) \;=\; \prod_{d=1}^{D} p_{\eta}(\tilde{y}_d \;-\; \boldsymbol{\mathcal{M}_d}(\boldsymbol{X})),$$

where there are D conditions (e.g., the time snapshots of measured streamflow) that are being used for inference. Typically, the marginal discrepancy random variable η is assumed to be Gaussian, essentially corresponding to the probabilistic counterpart of a least-square fit.

Text S5. Dimensional reduction of quantities of interest

There can be many quantities of interest (QoIs) in flood modeling, including flood depths, flow velocities, pressures at specific locations in the watershed of interest and at certain times. Given the spatiotemporal nature of these quantities, there may be strong correlations among QoIs in space or time. In this case, it can be beneficial to reduce the dimensionality of the number of

QoIs to estimate. This can be done by considering the output space as a random field and using the Karhunen-Loève (KL) decomposition [*Karhunen*, 1946]:

$$\mathcal{M}(x;\, \boldsymbol{X}) \;\approx\; \bar{\mathcal{M}}(x) + \sum_{d=1}^{D} \xi_d(\boldsymbol{X}) \sqrt{\mu_d}\varphi_d(\boldsymbol{x}), \qquad (3)$$

where $\mathcal{M}(x;\boldsymbol{X})$ provides the output from the model $\mathcal{M}$ given design conditions x (location, time, model output) and uncertain parameters $\boldsymbol{X}$. $\bar{\mathcal{M}}(x)$ is the mean of the random field, and the remainder is the decomposition of the variability in the random field, where $\xi_d(\boldsymbol{X})$ are Karhunen-Loève coefficients, and $\sqrt{\mu_d}\varphi_d(\mathrm{x})$ are KL modes scaled by the square root of the eigenvalues of the covariance matrix of the random field.

The benefit of this approach is that it reduces the number of surrogates (i.e., the initial number of QoIs) that need to be constructed to D KL coefficients. If we take the number of design conditions to be N_x, then often times $D \ll N_x$, and this approach has the opportunity to provide appreciable computational efficiency (see Text S10 for more details). As the dimension of the spatial domain grows, this type of decomposition becomes increasingly valuable in improving computational efficiency, allowing for modelers to have more flexibility in the types of model outputs that can be investigated in real-time.

Text S6. Rainfall and its uncertainty

Both measuring and forecasting rainfall for urban flood applications is a major challenge, due to the need for both high accuracy and high spatial and temporal resolution. Numerical weather prediction (NWP) models in particular struggle to adequately forecast extreme rainfall rates at

the correct location and time. Different observations and dataset characteristics can result in widely diverging estimates of rainfall for individual extreme storms. We illustrate this issue in Figure S5, which shows rainfall estimates for the period of analysis (0-18 UTC on 27 August 2017) from five sources: 1) a gage-corrected radar rainfall product (5-minute, 500 m x 500 m resolution; see details below); 2) Integrated Multi-satellitE Retrievals for GPM from NASA (IMERG 30-minute, 0.1^{o} resolution, *Huffman et al.* [2015]); 3) High-Resolution Rapid Refresh forecast product (HRRR; 1-hour, 3000 m resolution forecast issued as at the event onset; see *Benjamin et al.* [2016] for details); 4) rain gage time series from Weather Underground (Station ID: KTXHOUST1941, it is the nearest station to the case study; and 5) the Regional Spectral Model [*Nunes*, 2016]; 12 km x 12 km, 3-hour outputs, assimilating the 8-km NOAA's Climate Prediction Center Morphing precipitation [*Joyce et al.*, 2004], CMORPH, interpolated to a 12-km resolution grid, with initial and boundary conditions from the NCEP-DOE (National Centers for Environmental Prediction-Department of Energy) R2 global reanalysis, with integration starting time at 0 UTC on 15 August 2017. The differences among the products highlights the inherent issue of precipitation uncertainty. Figure S6 illustrates the instantaneous spatial rainfall distribution in the watershed from four precipitation sources, highlighting the considerable variability among sources, which will translate to substantial differences in the timing, extent, and magnitude of forecasted floods.

The gage-corrected radar rainfall dataset used in the study was generated based on a conversion of specific differential phase to rainfall rate, using dual-polarimetric information collected from the NEXRAD (Next-Generation Radar) KHGX installation in Houston, Texas. Five-minute accumulations from Harris County Flood Control District rain gage data were used to perform mean-field bias correction of the radar data.

Text S7. Mesh generation

The flowchart in Figure S7 outlines the process of mesh generation that accounts for the complexity of geometry of the urban environment. Specifically, ArcGIS Hydrology Package (Version 10.5.1. Redlands, CA: Environmental Systems Research Institute, Inc., 2019) hydrologic analysis tools are used to delineate the watershed area (46.9 km^2) using the 3 m resolution Digital Elevation Model (DEM) developed by the USGS. ArcGIS tool "Raster2TIN" is then used to develop Triangulated Irregular Network (TIN) from the delineated grid-based watershed. TIN is a representation of a surface with triangular facets that can capture terrain geometrical features with less number of nodes, as compared to the original raster-based description of the watershed.

To account for the effects of buildings on flood wave propagation, we next merge building footprint data as a set of irregularly spaced TIN nodes. As the original building footprint has a significant number of small-scale features, these are first simplified with the ArcGIS tool "Simplify Polygon". Figure S8 illustrates a small region to showcase this procedure. There are two major reasons for simplifying the building footprints: (1) to avoid the occurrence of very small triangles, targeting the minimum characteristic dimension of $\sim\mathcal{O}(10^0)$ m with an area < 0.1 km^2, as the smallest area of a computational cell constrains the time step in the finite-volume method implemented in OFM to simulate overland flow; and (2) to decrease the number of cells in the overall mesh representing the watershed, so as to reduce the computational burden of numerical simulation.

Another necessary element of mesh generation is the specification of channel and floodplain areas. Conventionally, a stream network can be extracted from the DEM based on flow

accumulation analysis. In a heavily urbanized region, the streamlines become discontinuous where bridges and roadways cross river channels. In order to guarantee the connectivity of the stream network, we incorporate additional stream vertices discretized from the continuous streamlines to the set of TIN nodes from the terrain and building footprint analysis. Finally, Delaunay triangulation [*Okabe et al.*, 2009] is carried out using the developed node set.

In summary, the TIN for the case study watershed results in 136,423 mesh nodes and 271,215 triangle cells. There are about 20,000 buildings included in the mesh, which take up approximately 30% of the total watershed area. All of the land cover is developed area: open space (13%), low intensity (32%), medium intensity (35%), and high intensity (21%). The land use information was downloaded from National Land Cover Database 2016.

Text S8. High-fidelity flood model

A coupled model integrating hydrologic, hydraulic, and morphologic processes, previously verified, and applied to real watersheds [*Kim et al.*, 2013; *Kim et al.*, 2012] was used in this research as the high-fidelity flood model $\mathcal{M}$ (Text S3). The hydrology module, the TIN - Based Real Time Integrated Basin Simulator (tRIBS), can simulate a range of hydrological processes such as canopy interception, evapotranspiration from bare soil and canopy, vertical and lateral moisture fluxes in the subsurface, and various runoff generation mechanisms (i.e., saturation excess, infiltration excess, perched subsurface stormflow, and groundwater exfiltration), given the spatial inputs of meteorological data, topography, landuse, and soil type [*Ivanov et al.*, 2008; *Ivanov et al.*, 2004]. Accounting for these hydrologic processes offers water sources/sinks to a model simulating the hydrodynamics of overland flow – Overland Flow Model, OFM [*Kim et*

al., 2012], using physically modeled wave speeds within the domain of arbitrary geometric configuration.

In the case study considered, the computed flow states did not influence runoff generation, implying that water flowing on surface is not allowed to re-infiltrate. Furthermore, it was assumed that no water loss took place due to infiltration or stormwater management infrastructure. This assumption is justified for the August 27, 2017 event, which followed a series of antecedent storms that likely filled the soil water and stormwater storage capacities very quickly after the time of event onset. The inclusion of both types of rainfall losses in the high-fidelity model tRIBS-OFM is possible and should be used in assessments of less extreme events. It would however result in an additional computational demand, further emphasizing the need for the developed flood-forecasting framework.

The principal elements of the hydrodynamic model OFM are outlined below.

The OFM model solves the 2D Saint-Venant equations (i.e., the shallow water equations), which are derived by depth-integrating the Navier-Stokes equations [*Liggett*, 1968]. The governing equations consist of a continuity equation and two momentum equations for two perpendicular horizontal directions:

$$\frac{\partial U}{\partial t} + \frac{\partial E}{\partial x} + \frac{\partial G}{\partial y} = S\,, \qquad (4)$$

where U is vector of flow variables, E and G are the flux terms in x and y direction, respectively, and S is the source vector [*Kim et al.*, 2012]:

$$U = \begin{bmatrix} h \\ uh \\ vh \end{bmatrix}, E = \begin{bmatrix} uh \\ u^2h + \frac{1}{2}gh^2 \\ uvh \end{bmatrix}, G = \begin{bmatrix} vh \\ uvh \\ v^2h + \frac{1}{2}gh^2 \end{bmatrix}, \qquad (5)$$

$$S = \begin{bmatrix} i \\ -gh\frac{\partial z_b}{\partial x} - C_D u\sqrt{u^2+v^2} \\ -gh\frac{\partial z_b}{\partial y} - C_D \sqrt{u^2+v^2} \end{bmatrix}, \qquad (6)$$

where h represents flow depth, u, v are the flow velocities in x-axis and y-axis direction in the Cartesian system of coordinates, g is the gravitational acceleration constant, i is the net runoff production rate, z_b is the bed elevation, and $C_D = gn^2h^{-1/3}$ is the surface drag coefficient, where n is Manning roughness coefficient (assumed to be spatially uniform at value of 0.015 in the case study watershed because the sensitivity to rainfall input overrides the sensitivity to roughness). Detailed descriptions of the numerical solution with the finite-volume method can be found in *Kim et al.* [2012].

The solution is valid for most kinds of overland flow conditions. For numerical stability, in the time-explicit finite volume method the time step is constrained by the mesh's smallest cell area, which introduces another challenge in the forecast with the overland model. For example, the typical time step is ~$\mathcal{O}(10^{-1})$ sec for a high-resolution application (e.g., cell area is ~$\mathcal{O}(10^{1})$ m). Time-implicit schemes for solving shallow water equations do not impose hard constraints on time stepping and require less solution steps, offering computational savings [*Fernández-Pato et al.*, 2018]. The overall computational need however depends on the number of iterations within each time step and for a transient flow situation, such as flood wave motion, the demand is

higher. This limits the possible gains from an implementation of this numerical scheme to less than one order of magnitude.

Text S9. Rainfall input for pretraining and forecasting

The approach requires pretraining simulations using the high-fidelity flood-resolving hydrologic and hydrodynamic model of *Kim et al.* [2012] combined with time series of input rainfall. Outputs from these high-fidelity simulations are used to train reduced-order models that will 'learn' the underlying physics (see Text S3). As the specifics of a rainfall event to be forecasted in future cannot be known during the pretraining phase (the proposed framework postulates that it needs to be during a pre-storm period), it must be determined what kind of rainfall input should be supplied to the high-fidelity model. While these inputs could take the form of previous observed storms or hypothetical "synthetic" events, a more powerful demonstration of the efficacy of the proposed approach is if pretraining simulations are fully "ignorant" to circumstances of a flooding event for which a forecast would have to be issued. As a result, we developed an ensemble set of 18-hour long rainfall inputs, each represented as a series of uncorrelated pulses randomly drawn from the Uniform distribution between 0 and 50 [mm/hour]. These inputs were spatially uniform, chosen due to the fairly modest size of the study watershed. The addition of realistic spatiotemporal evolution of storm structure using stochastic representations of precipitation process [*Fatichi et al.*, 2013; *Peleg et al.*, 2017] is a logical extension of the framework. It would provide rainfall series for pretraining simulations that have a structure similar to observations.

After having completed the pretraining and in order to emulate a real-time operational application of the framework, we use an 18-hour of the High Resolution Rapid Refresh (HRRR) product provided by NOAA at the event onset. The quantitative precipitation forecast product one of many precipitation estimates that NOAA released to water resources and forecasting professionals prior to or following this forecast. The HRRR rainfall forecast scenario was used to construct a Gaussian process (GP) stochastic model [*Rasmussen and Williams*, 2006] of the event precipitation process. The GP model was parameterized as follows:

$$f(t) \sim GP(f(t) \mid m(t), k(t,t'), \qquad (7)$$

where $f(t)$ represents precipitation intensity at hour t, $m(t)$ is the mean precipitation intensity at hour t, and $k(t,t')$ is the covariance function between time t and t'. In this study, squared-exponential kernel was used to represent similarity of precipitation intensity between times t and t', i.e., how close the corresponding precipitation rates are

$$k(t,t') = \sigma^2 exp(-(t-t')^2 / 2l^2), \qquad (8)$$

where σ^2 is the variance, and l is the length scale. Specifically, we used σ^2 = 467 and l = 0.74 when fitting Gaussian Process to the HRRR precipitation forecast.

By sampling from the GM model, a set of rainfall realizations was generated for the forecast period and used as inputs into the surrogate models for the real-time flood forecasting and uncertainty quantification demonstrations.

Text S10. Computational efficiency of surrogate models

Due to its polynomial form, the reduced-order surrogate model is much more computationally efficient than the high-fidelity model. About 2.84 CPU seconds (Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz) was sufficient to run all of the 127,026 constructed surrogate models (i.e., 7,057 locations ×18 hours). Thus, *Core Seconds per Simulated Day - Normalized* (CSSD-N, Text S2) metric for polynomial chaos surrogate (PCS) can be computed as $\eta_{PCS} = \frac{2.84}{7507} \times \frac{24}{18} = 0.0005$, or $\mathcal{O}(10^{-4})$ CPU seconds to carry out 24-hour simulation for a single location (i.e., mesh node of the high-fidelity model). This translates to a computation time reduction of 2-4 orders of magnitude compared with the high-fidelity flood model (the corresponding range for η_{FHFM} is assessed as $\mathcal{O}(10^{-2})$ - $\mathcal{O}(10^{0})$), which is sufficiently fast for forward runs in real-time flood-forecasting.

The number of simulations needed to carry out uncertainty analysis is still substantial, however: at the least 127,026 ×$\mathcal{O}(10^{1}) \rightarrow \mathcal{O}(10^{6})$ of basic algebraic operations are embedded in polynomial expansion series per single forward run. They may pose computational challenges, especially when these are run serially on a typical desktop or laptop. For example, carrying out 10,000 simulations for each QoI to quantify the uncertainty stemming from precipitation input can take about eight hours; even more so, if the number of selected QoIs is orders of magnitude higher. However, if the dimensionality of QoIs can be reduced significantly because of their high spatial correlation (Text S5 “Dimensional reduction of quantities of interest”), there is no need to formally carry out uncertainty quantification for all QoIs. In this study, for example, using the first three dominant eigenvalues retains 99% of the original training sets for flooding depth at a given time instant (Figure S9). Therefore, the use of KL decomposition reduces the number of

surrogates that needed to be constructed from 127,026 to 54. This is achieved because a surrogate model was constructed at each 7,057 locations for each of 18 hours; by using the KL decomposition flooding depth spatial field can be reduced by using three KL modes for each of the 18 hours (7,057 × 18 vs. 3 × 18). As a result, this increases the numerical efficiency by additional three orders of magnitude: η_{PCS}~ $\mathcal{O}(10^{-7})$. This guarantees the efficiency of uncertainty quantification using the proposed framework relying on surrogate models. As the dimension of the spatial domain grows, this type of decomposition becomes increasingly valuable in improving computational efficiency, allowing for modelers to have more flexibility in the types of model outputs that can be investigated in real-time.

We note that the additional computational effort of surrogate training based on outputs of high-fidelity model represents only a small fraction of the time required by the high-fidelity model (i.e., 52 hours in Text S2). Using the Bayesian regression [*Sargsyan et al.*, 2014], it takes 2.03 hours to construct all of the 127,026 surrogates, or ~0.0575 or $\mathcal{O}(10^{-2})$ CPU seconds per surrogate. The use of KL decomposition (see the paragraph above) reduces the overall computational effort further: it takes 414 seconds to carry out the KL decomposition analysis to reduce the number of surrogates that needed to be constructed from 127,026 to 54; the subsequent Bayesian regression to estimate PCE coefficients takes 37.8 seconds. This results in $\frac{[414+37.8]}{127{,}026} = 0.00356$ or $\mathcal{O}(10^{-3})$ CPU seconds per surrogate (note that we assume that 54 represent all 127,062). As illustrated above, this permits the increase of the overall numerical efficiency of surrogates by three orders of magnitude (i.e., η_{PCS}~ $\mathcal{O}(10^{-7})$).

The reallocation of intense simulations with the high-fidelity model from the flooding periods to the time interval between them effectively replaces the strain of computational burden in real-

time with a data storage problem. However, the actual amount of storage is likely to remain modest: in the case study considered, 3 gigabytes of outputs from the high-fidelity model are used to train 127,026 surrogate models, and their multi-index arrays total in volume of ~1.3 gigabytes. Normalizing the latter value by the total number of QoIs yields 0.01 megabytes per QoI, a trivial magnitude, given the low cost of modern storage systems. Since the QoI dimensionality can be further reduced using the Karhunen- Loève decomposition, the storage of multi-index arrays representing surrogate models is only ~0.5 megabytes.

Table S1: U.S. flood-related fatalities from 1995 to 2017. Data were downloaded from https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/. Some of the summaries are no longer accessible at the NOAA website but the authors have the entire record used to develop Table S1 available.

Year	Total	In Vehicle	In Water	Flash Flood	River Flood
1995	80	39	35	60	20
1996	131	79	31	94	37
1997	118	46	45	86	32
1998	136	75	25	118	18
1999	68	26	29	60	8
2000	38	24	7	30	8
2001	48	24	12	35	13
2002	49	28	11	38	11
2003	86	39	21	67	18
2004	82	45	25	58	24
2005	43	18	19	28	15
2006	76	32	34	59	17
2007	87	50	21	70	17
2008	82	39	23	58	24
2009	56	33	10	33	23
2010	103	45	22	67	36
2011	113	68	26	69	44
2012	29	10	11	19	10
2013	82	37	17	60	22
2014	40	16	14	29	9
2015	187	112	24	129	45
2016	126	58	34	86	40
2017	136	40	52	N/A	N/A
Average	86.8	42.7	23.8	61.5	22.3
Percentage	100%	49.3%	27.5%	70.9%	25.7%

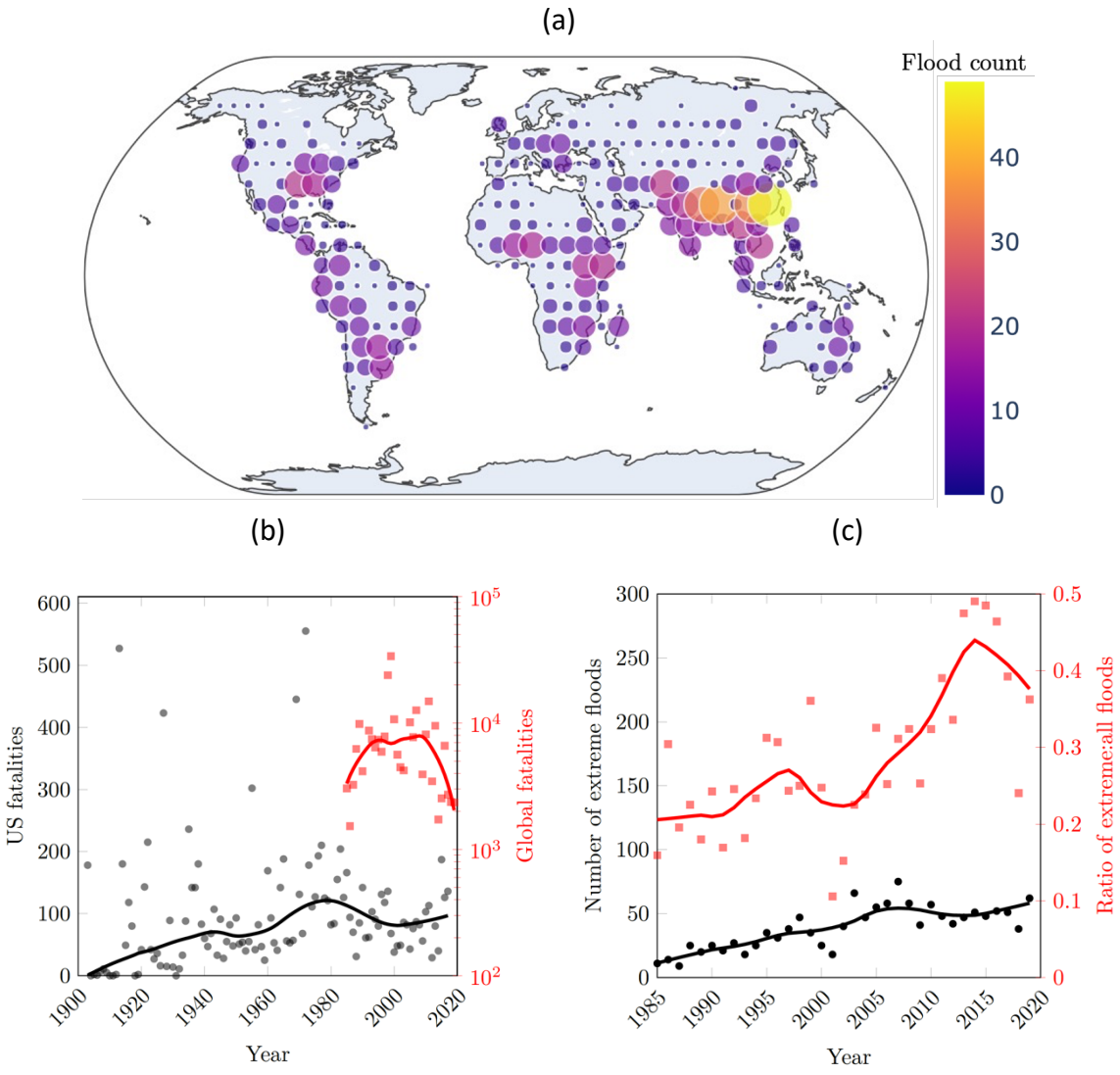

Figure S1. Floods and fatalities. (a) Count of extreme floods for 1985-2019 on a 7.5°×7.5° latitude - longitude grid. (b) Annual fatalities from flooding globally (red) and in the U.S. (black). (c) The total number of annual extreme floods globally (black) and the ratio of extreme floods to all floods (red). The data sources and the definition of extreme floods are described in Text S1.

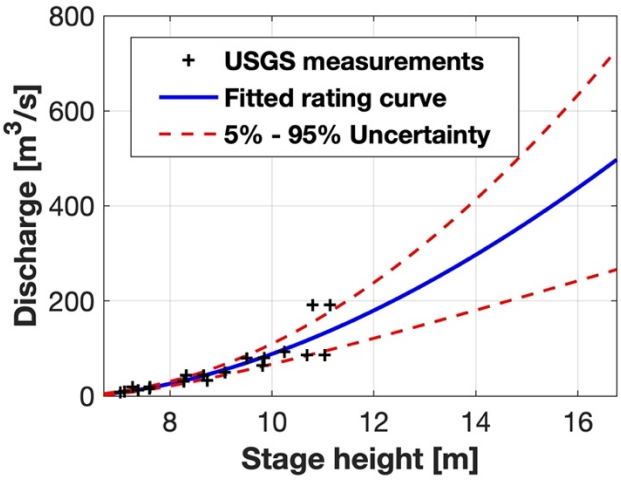

Figure S2. Stage-discharge relationship and streamflow uncertainty. River stage – discharge data (i.e., the rating curve) collected by the U.S. Geological Survey at gage 08074540. The red dashed line indicates the 5%-95% regional of uncertainty associated with the stage-discharge relationship estimated using the "ISO/WMO" uncertainty assessment method [*Kiang et al.*, 2018].

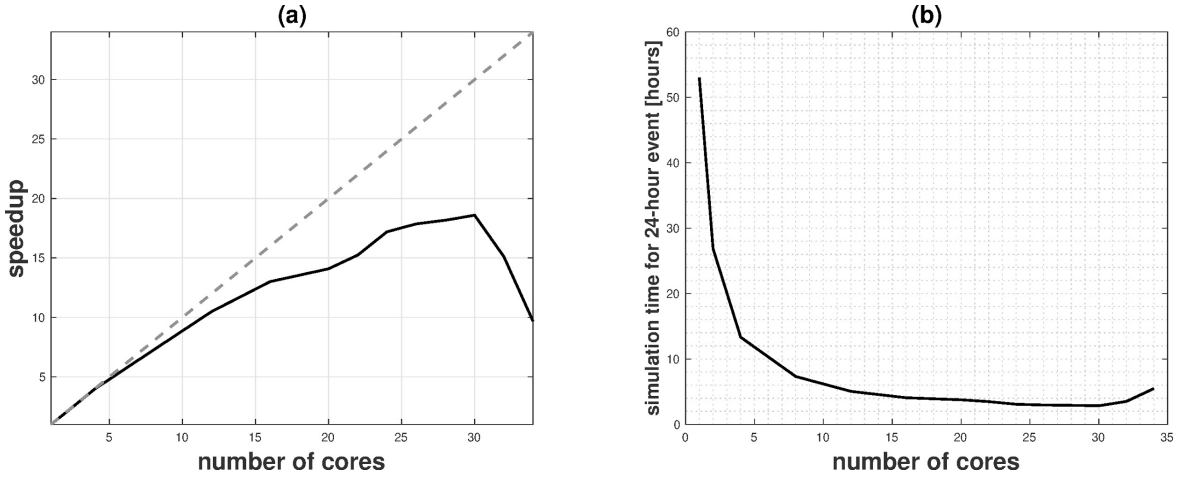

Figure S3. Computational performance of parallelized Overland Flow Model (OFM). The model is by Kim *et al*. (2012). (a) Speedup factor as a function of number of CPU cores, with the dashed line illustrating a perfect scalability. The factor is defined as the ratio of model execution time in serial mode to that in parallelized mode. (b) OFM simulation time for 24-hour event using the case study watershed.

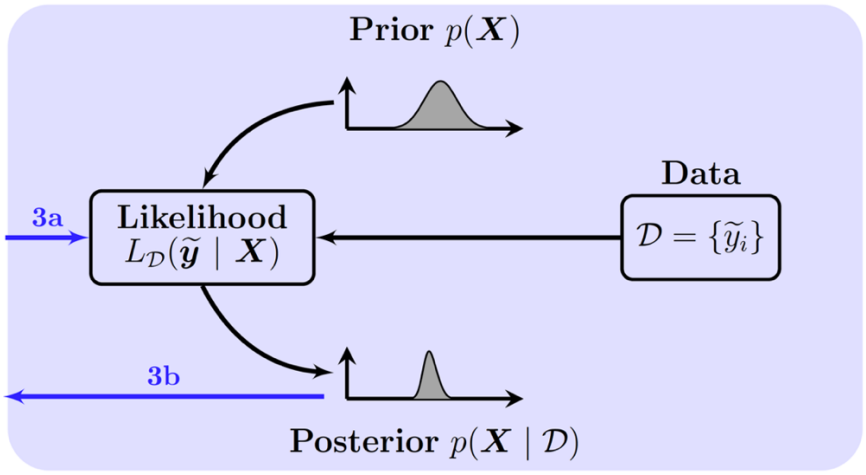

Figure S4. Workflow of the inference portion of the uncertainty quantification workflow. Steps 3a and 3b refer to the data flows present in Figure 1 of the main narrative, where 3a is the flow of outputs of QoIs from the surrogate model into a likelihood function to compare the value from the surrogate and observed data. This inference is used to obtain posterior distribution for $\boldsymbol{X}$, trained on the observational data $\tilde{\boldsymbol{y}}$. This posterior can then be sent back to the surrogate in 3b in order to reduce the uncertainty in the inputs that are evaluated through the surrogate model.

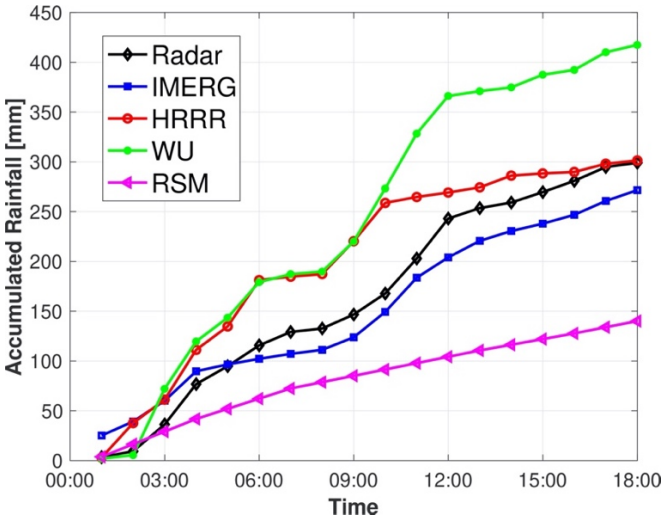

Figure S5. Cumulative, area-averaged rainfall for the case study watershed on August 27, 2017. Data (time axis uses the Greenwich Mean Time): the high-resolution gage-adjusted radar product ("radar", black diamonds), NASA IMERG product ("IMERG", blue squares), NOAA High Resolution Rapid Refresh ("HRRR", red circles), Weather Underground ("WU", green circles) and simulation with the Regional Spectral Model ("RSM", magenta triangles).

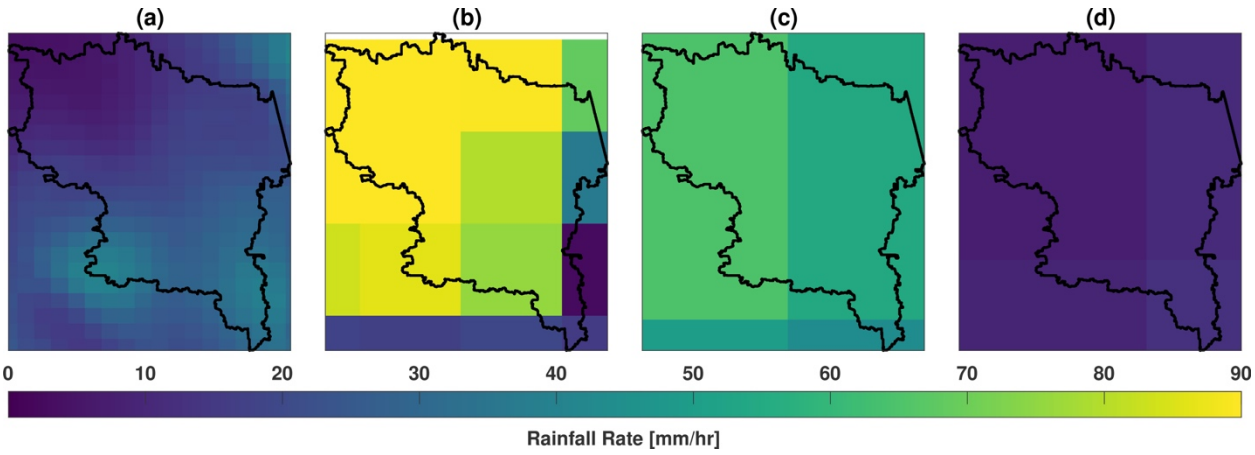

Figure S6. Spatial distribution of rainfall. A snapshot is at 4:30 UTC on August 27 2017 from (a) gage-adjusted radar rainfall product, (b) NASA IMERG satellite-based product, (c) NOAA HRRR forecast product, and (d) RSM simulation.

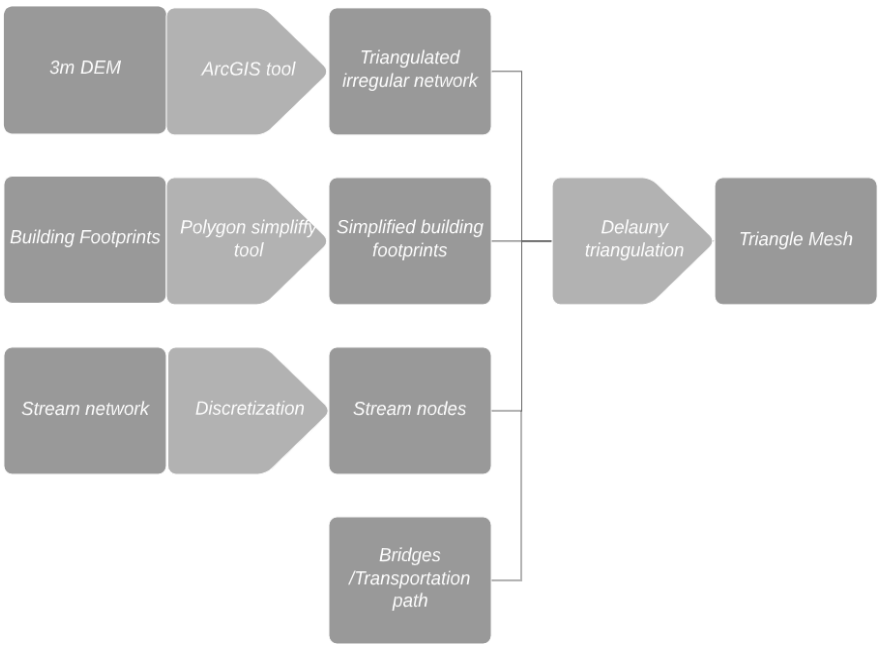

Figure S7. Flowchart for mesh generation process for an urban domain. Rectangles represent elements of the mesh at various stages; pointed rectangles represent a process / tool by which the various mesh elements are modified.

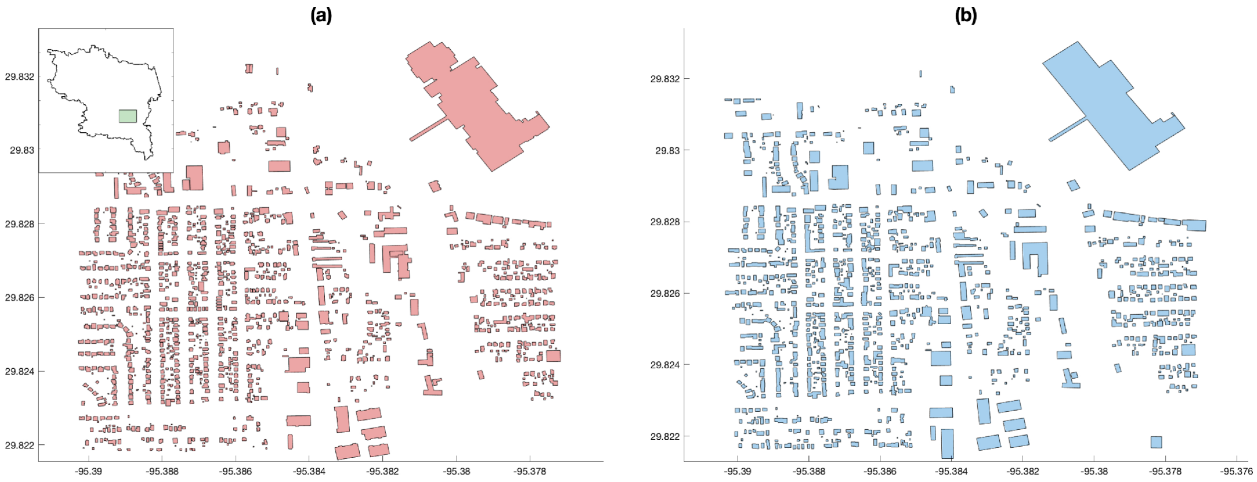

Figure S8. An example of building footprint layout simplification. Subplot (a) illustrates the original building footprint layout (the inset shows the location of the region within the watershed area as the green box), and subplot (b) shows the simplified building shapes.

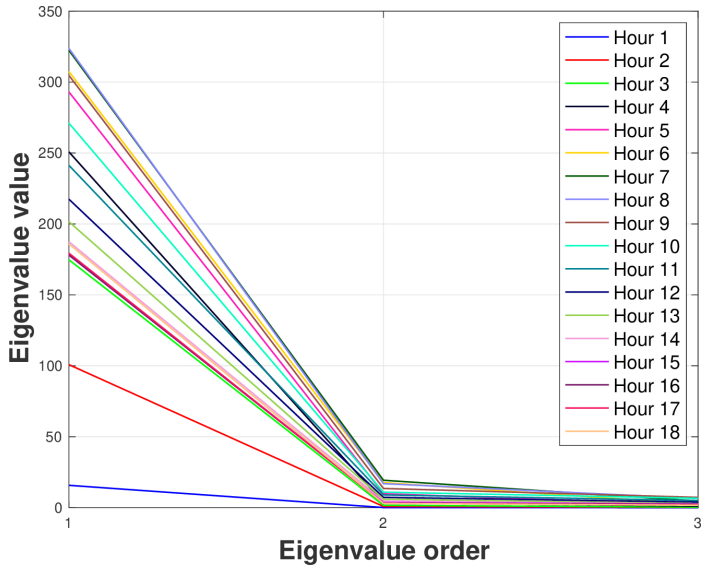

Figure S9. An illustration of dimensionality reduction for simulated inundation depth. The first three eigenvalues from the training sets of Region 2 at all the 18 hours. Specifically, the eigenvalues are computed from 4,476 × 1,000 matrix, where 4,476 is the number of cells in the case study watershed subareas and 1,000 is the number of training simulations.

References:

Abe, S. (2010), *Support Vector Machines for Pattern Classification*, Springer, London.

Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger (2013), GloFAS-global ensemble streamflow forecasting and flood early warning, *Hydrology and Earth System Sciences*, *17*(3), 1161.

Artichowicz, W., and D. Gasiorowski (2019), Computationally efficient solution of a 2D diffusive wave equation used for flood inundation problems, *Water*, *11*(10), doi:10.3390/w11102195.

Balaji, V., E. Maisonnave, N. Zadeh, B. N. Lawrence, J. Biercamp, U. Fladrich, G. Aloisio, R. Benson, A. Caubel, J. Durachta, M. A. Foujols, G. Lister, S. Mocavero, S. Underwood, and G. Wright (2017), CPMIP: measurements of real computational performance of Earth system models in CMIP6, *Geoscientific Model Development*, *10*(1), 19-34, doi:10.5194/gmd-10-19-2017.

Benjamin, S. G., S. S. Weygandt, J. M. Brown, M. Hu, C. R. Alexander, T. G. Smirnova, J. B. Olson, E. P. James, D. C. Dowell, G. A. Grell, H. Lin, S. E. Peckham, T. L. Smith, W. R. Moninger, J. S. Kenyon, and G. S. Manikin (2016), A North American hourly assimilation and model forecast cycle: the rapid refresh, *Monthly Weather Review*, *144*(4), 1669-1694, doi:10.1175/mwr-d-15-0242.1.

Brakenridge, G. R. (2016), Global active archive of large flood events. *Rep.*

Cleveland, W. S., and S. J. Devlin (1988), Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association*, *83*(403), 596-610.

Dwelle, M. C., J. Kim, K. Sargsyan, and V. Y. Ivanov (2019), Streamflow, stomata, and soil pits: Sources of inference for complex models with fast, robust uncertainty quantification, *Advances in Water Resources*, *125*, 13-31, doi:https://doi.org/10.1016/j.advwatres.2019.01.002.

Fatichi, S., V. Y. Ivanov, and E. Caporali (2013), Assessment of a stochastic downscaling methodology in generating an ensemble of hourly future climate time series, *Climate Dynamics*, *40*(7-8), 1841-1861, doi:10.1007/s00382-012-1627-2.

FEMA (2018), FEMA - Harvey Flood Depths Grid, HydroShare, in *Harvey Flood Depths Grid, HydroShare*, edited.

Fernández-Pato, J., M. Morales-Hernández, and P. García-Navarro (2018), Implicit finite volume simulation of 2D shallow water flows in flexible meshes, *Computer Methods in Applied Mechanics and Engineering*, *328*, 1-25.

Huffman, G. J., D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, P. Xie, and S.-H. Yoo (2015), NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG), *Algorithm Theoretical Basis Document (ATBD) Version*, *4*, 26 pp.

Ivanov, V. Y., R. L. Bras, and E. R. Vivoni (2008), Vegetation-hydrology dynamics in complex terrain of semiarid areas: 1. A mechanistic approach to modeling dynamic feedbacks, *Water Resources Research*, *44*(3), doi:10.1029/2006wr005588.

Ivanov, V. Y., E. R. Vivoni, R. L. Bras, and D. Entekhabi (2004), Catchment hydrologic response with a fully distributed triangulated irregular network model, *Water Resources Research*, *40*(11), doi:10.1029/2004wr003218.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press: Cambridge.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie (2004), CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution, *Journal of Hydrometeorology*, *5*(3), 487-503.

Karhunen, K. (1946), Zur spektraltheorie stochastischer prozesse, *Annales Academiae Scientiarum Fennicae. Mathematica-Physica*, *34*.

Kennedy, M. C., and A. O'Hagan (2000), Predicting the output from a complex computer code when fast approximations are available, *Biometrika*, *87*(1), 1-13.

Kiang, J. E., C. Gazoorian, H. McMillan, G. Coxon, J. Le Coz, I. K. Westerberg, A. Belleville, D. Sevrez, A. E. Sikorska, and A. Petersen-Øverleir (2018), A comparison of methods for streamflow uncertainty estimation, *Water Resources Research*, *54*(10), 7149-7176.

Kim, J., V. Y. Ivanov, and N. D. Katopodes (2013), Modeling erosion and sedimentation coupled with hydrological and overland flow processes at the watershed scale, *Water Resources Research*, *49*, 5134-5154, doi:doi:10.1002/wrcr.20373.

Kim, J., A. Warnock, V. Y. Ivanov, and N. D. Katopodes (2012), Coupled modeling of hydrologic and hydrodynamic processes including overland and channel flow, *Advances in Water Resources*, *37*, 104-126, doi:https://doi.org/10.1016/j.advwatres.2011.11.009.

Le Maître, O., and O. M. Knio (2010), *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*, Springer Science & Business Media.

Liggett, J. A. (1968), Mathematical flow determination in open channels, *Journal of the Engineering Mechanics Division - ASCE*, *94*(EM4), 947-963.

Marzouk, Y., and D. Xiu (2009), A stochastic collocation approach to Bayesian inference in inverse problems, *Communications in Computational Physics*, *6*(4), 826-847.

Marzouk, Y. M., H. N. Najm, and L. A. Rahn (2007), Stochastic spectral methods for efficient Bayesian solution of inverse problems, *Journal of Computational Physics*, *224*(2), 560-586, doi:10.1016/j.jcp.2006.10.010.

Najm, H. N. (2009), Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics, *Annual Review of Fluid Mechanics*, *41*, 35-52.

Neal, J. C., T. J. Fewtrell, P. D. Bates, and N. G. Wright (2010), A comparison of three parallelisation methods for 2D flood inundation models, *Environmental Modelling & Software*, *25*(4), 398-411, doi:https://doi.org/10.1016/j.envsoft.2009.11.007.

Noh, S. J., J.-H. Lee, S. Lee, and D.-J. Seo (2019), Retrospective dynamic inundation mapping of hurricane Harvey flooding in the Houston metropolitan area using high-resolution modeling and high-performance computing, *Water*, *11*(3), 597.

Nunes, A. M. B. (2016), Satellite-enhanced dynamical downscaling for the analysis of extreme events, *Journal of Geophysical Research: Atmospheres*, *121*(18), 10-617.

Okabe, A., B. Boots, K. Sugihara, and S. N. Chiu (2009), *Spatial tessellations: concepts and applications of Voronoi diagrams*, John Wiley & Sons.

Peleg, N., S. Fatichi, A. Paschalis, P. Molnar, and P. Burlando (2017), An advanced stochastic weather generator for simulating 2-D high-resolution climate variables, *Journal of Advances in Modeling Earth Systems*, *9*(3), 1595-1627, doi:10.1002/2016MS000854.

Rasmussen, C. E., and C. K. I. Williams (2006), Gaussian processes for machine learning, in *Gaussian processes for machine learning*, edited, the MIT Press.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Sanders, B. F., and J. E. Schubert (2019), PRIMo: Parallel raster inundation model, *Advances in Water Resources*, *126*, 79-95.

Sargsyan, K., H. N. Najm, and R. Ghanem (2015), On the statistical calibration of physical models. *International Journal of Chemical Kinetics*, 47 (4), 246–276. https://doi.org/10.1002/kin.20906.

Sargsyan, K., C. Safta, H. N. Najm, B. J. Debusschere, D. M. Ricciuto, and P. Thornton (2014), Dimensionality reduction for complex models via Bayesian compressive sensing *International Journal for Uncertainty Quantification*, *4*, 63-93, doi:10.1615/Int.J.UncertaintyQuantification.2013006821.

Tarantola, A. (2005), Inverse problem theory and methods for model parameter estimation, *vol 89,* Philadelphia, PA: SIAM, doi:10.1137/1.9780898717921.

Wiener, N. (1938), The homogeneous chaos, *American Journal of Mathematics*, *60*(4), 897-936.

Xiu, D., and G. E. Karniadakis (2002), The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM Journal on Scientific Computing*, *24*(2), 619-644.

Xiu, D., and G. E. Karniadakis (2003), Modeling uncertainty in flow simulations via generalized polynomial chaos, *Journal of Computational Physics*, *187*(1), 137-167.

Xiu, D., and D. M. Tartakovsky (2004), Uncertainty quantification for flow in highly heterogeneous porous media, in *Developments in Water Science*, edited by M. W. F. Cass T. Miller, William G. Gray and George F. Pinder, pp. 695-703, Elsevier.

Xu, D. (2020), Addressing uncertainty in understanding hydroclimate, hydrology and hydraulics across scales, Ph.D. dissertation, University of Michigan.