Discrepency and Mismatch between the ParFlow output and Observed data from USGS

immediate

1. Resolution

- **USGS** gage coordinates are given in latitude/longitude (geographic coordinates).
- **ParFlow** outputs are in structured grid space (i,j) at a specific resolution. might fall between grid cells

2. Potential Problem when resolution differ

- **Mismatch Spatial locations:** A USGS gage points might fall between grid cells or on the edge of the domain. Thus, this could be a source of uncertainty, as it may be assigned to a stream cell that it does not share the same characteristics of the original location of the gage station on the real world.
- **Gage might fall exactly inside a stream cell:** Projecting USGS on the CONUS 2 domain , which is structured on a high resolution grid, necessitates spatial approximation. We are making an assumption by assigned one single stream value to a whole single cell that represents a 1000 m in high x 1000 m length. Moreover, these cell may not share the same characteristics of the original location of the gage station which could be a source of uncertainty .

3. Other sources of Uncertainties

In addition to resolution - driven spatial mismatch, several other sources of uncertainty may contribute to discrepancies between streamflow outputs generated by ParFlow and these observed at USGS gage stations.

- **Land Cover Heterogeneity:** It plays a critical role in surface roughness and infiltration behavior
- **ParFlow** often applies a uniform Manning's value per grid cell potentially overlooking sub-grid scale.
- **Terrain Slope:** It is derived from Digital Elevation Model (DEM) can also affect flow directions and velocities calculations . While USGS gage measurements are considered highly reliable but they are still subject to measurements error:
- **Anthropogenic activities:** It is important to count for anthropogenic activities such as dam operations, withraw of water from river and other streams

4. How these factors affect the training of SBI to capture the true Manning's value of the observed data from USGS

- $x_{baseline}$: Simulated Streamflow from **ParFlow**
- $x_{observed}$: Observed streamflow from **USGS**

The problem is that we are training an SBI model (e.g: SNPE, SNL) to learn p(θ|x)
where:
θ: Manning's value
x: streamflow observation
But in our case, we are training inputs x come from ParFlow- simulated streamflow, which is affected by projection and spatial misalignment and other source of uncertainties.
we are assuming that:

$$x_{\text{observed}} \sim x_{\text{model}} \mid \theta$$

Thus, this leads to:

- **Posterior misspecification:** $x_{\text{model}} \mid \theta$ is biased because training data does not match reality.
- **Poor Calibration: Manning's value** are tuned to fit the model not the observed data.

5. Resolution

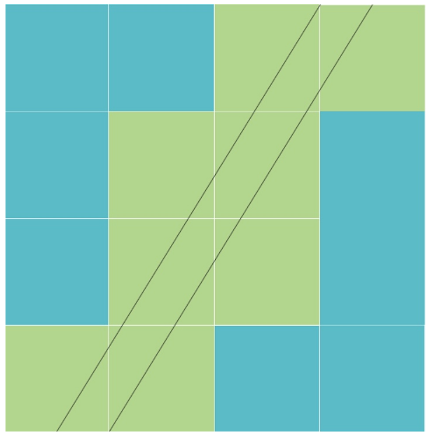

Figure 1. This is your caption.

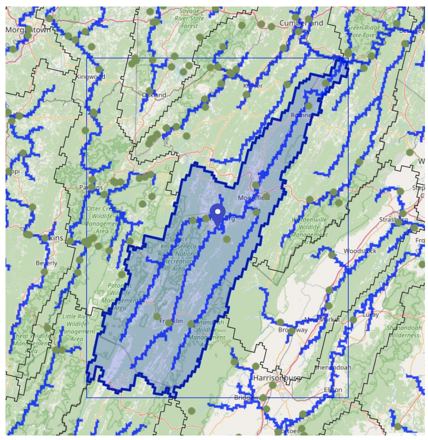

Figure 2. This is your caption.

6. Approach to address misalignment

We need to include e_i in our model:
First we know that:

$$f(\theta, t_i) = s_i$$

where:
$\theta\ :\ model parameters(Manning's value)$
$t_i\ :\ time$
$s_i\ :\ Simulated streamflow$

So, we can say that **ParFlow** produces streamflow outputs for each grid cell as a function of physical parameters and time.

The error (e_i) is defined as the difference between the streamflow observed data (at a specific point) and the model predicted (at a grid cell).

$$e_i = d_i - s_i$$

where:

d_i : observed data (e.g from USGS at a point location)
$s_i(\theta)$: simulated streamflow from PatFlow at a cell
This will address on the point versus cell mismatch.

We seek to determine the probability distribution of error.

$$p(e)$$

We know that:

$$p(e) = p(point) * p(cell \mid \theta)$$

We are suggesting a convolution between two probability distribution and it is meant to capture how the uncertainty in both the observation(point) and the simulation(cell) contributes to the overall uncertainty in the error e (the difference between simulated and observed streamflow).

7. Explanation of Convolution

Since we are comparing two random variables.

- The observation d $\sim p(POINT)$
- The simulation $s(\theta) \sim p(cell \mid \theta)$

It means that the total error is affected by :

- The inaccuracy or noise in your gage observation
- The inaccuracy of **ParFlow** simulation at grid cells (grid size, terrain misrepresentation, wrong Manning's value, and numerical error ...)

8. Error Distribution

We are assuming the following:

- $d \sim N(Q, \sigma^2_{\text{model}})$: The gage has a normal distribution centered around true flow Q, with some observation error.
- $s(\theta) \sim N(Q, \sigma^2_{\text{observed}})$: The model simulates flow with its own uncertainty.

$$e = d - s(\theta) \sim N(0, \sigma^2_{\text{observed}} + \sigma^2_{\text{model}})$$