difference between test set and flow histograms, including flow uncertainty (#samples = 10k)