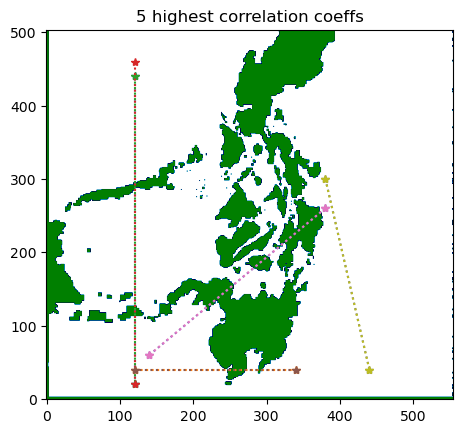
**Written Report – 6.419x Module 5** (Environmental Data) **Name**: DavidC\_

**Problem 2:** [10 bonus pts] *Identifying long-range correlations*.

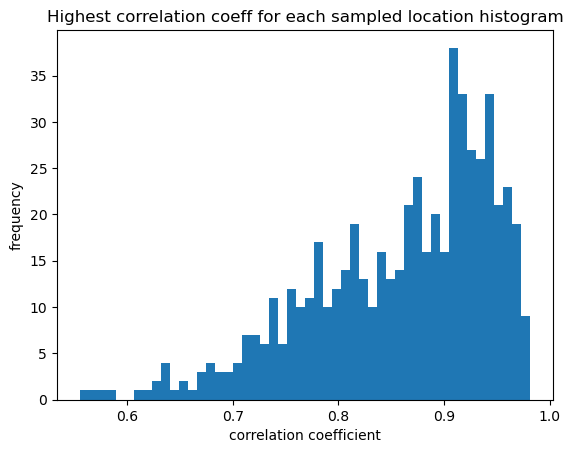
The plot below shows the 5 highest correlation coefficients (between pairs of locations) in Flow data.



This analysis was obtained by:

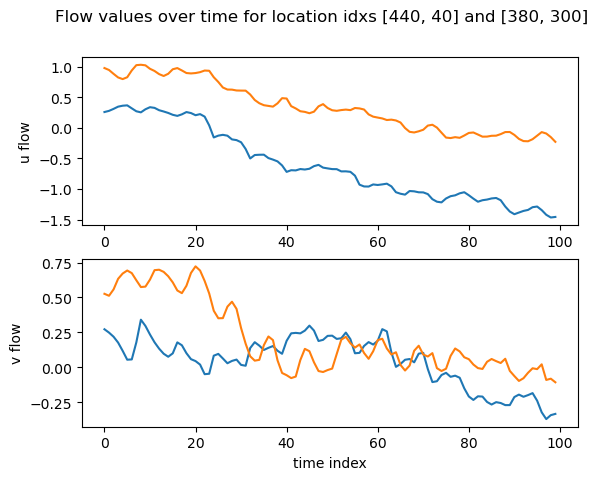
* Sampling data instead of checking every pair of 504x505 locations.
  + Sampled every 20th location: indices [1,20,40,60,...]
* Only check correlation coefficient if distance between locations was > 600 KM (200 indices) apart to avoid nearby locations.
* Calculate u and v component correlations separately and take the maximum as the highest for that location.

The highest correlation coefficient for each sampled location was mostly around 0.9 - 1.0, as shown in the histogram below:



Since many correlation coefficients are bunched near 1.0, I decided to show the highest 5 instead of just one.

The highest correlation coefficient occurred between locations (x index, y index): [440,40] and [380,300]. The flow data over time for that location pair is shown below. One can see high correlation in the u component.



**Problem 3.a.** *Provide simulation*

* ***(3 points)****: Provides an explanation of the simulation algorithm, with equations for the evolution of the particle trajectory.*
* ***(2 points)****: Provides a plot of the initial state of the simulation.*
* ***(3 points)****: Provides two plots of intermediate states of the simulation.*
* ***(2 points)****: Provides a plot of the final state of the simulation.*

Simulation algorithm:

for each increment of time (dt):

calculate a point's next location

x(t) is location of particle in x dimension

y(t) is location of particle in y dimension

u(i,t) is u\_flow value for index i and time t

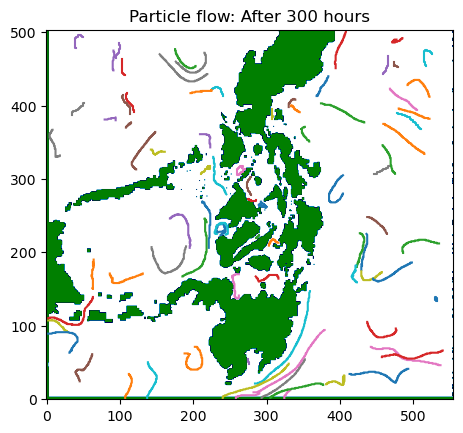
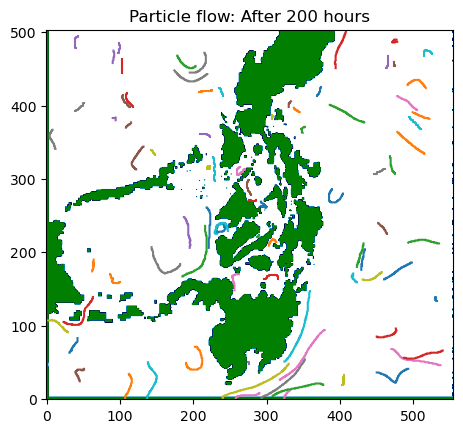
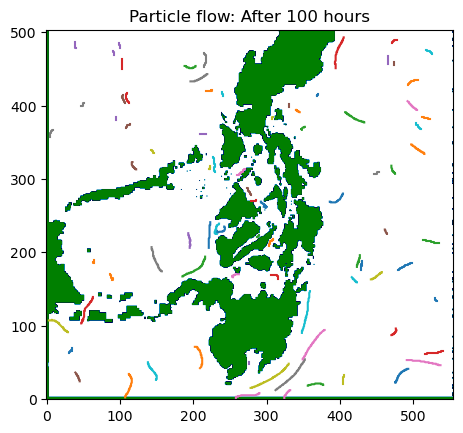
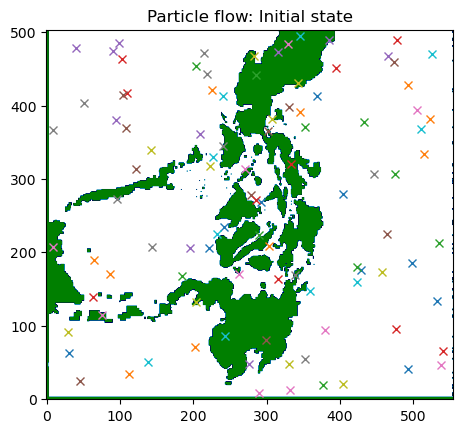
v(i,t) is u\_flow value for index i and time t

i = flow\_index(x(t)) // gets flow index for x km location

j = flow\_index(y(t)) // gets flow index for y km location

x(t+dt) = x(t) + u(i,t)

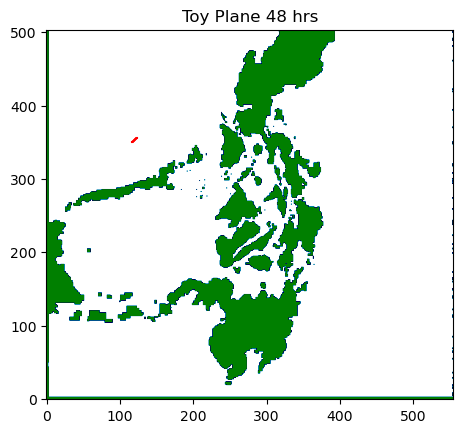
y(t+dt) = y(t) + v(j,t)

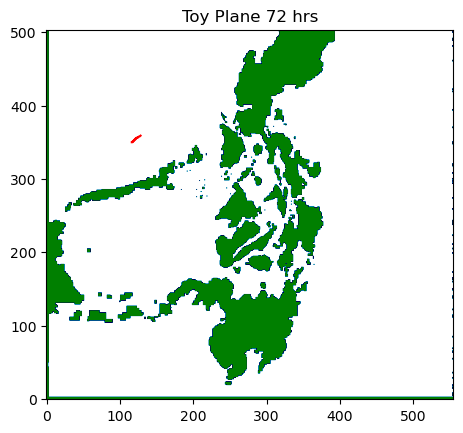


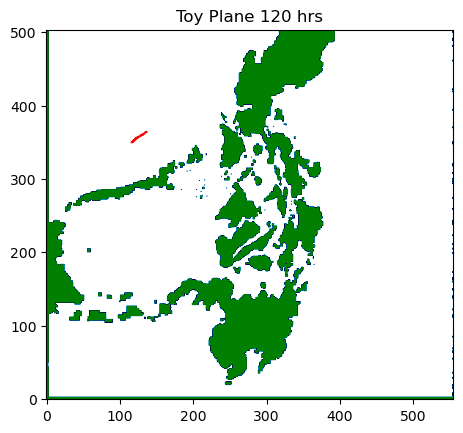
**Problem 3.b**: *Toy Plane crash simulation*

* ***(3 points)****: Provides plots showing the state of the simulation at the times: hrs, hrs, hrs. (Three plots required.)*
* ***(3 points)****: Two or more additional choices of the variances were tried, and three plots of the state of the simulation at the above three times are provided. (Six additional plots required.)*
* ***(4 points)****: Comments on where one should concentrate search activities based on the observed results.*

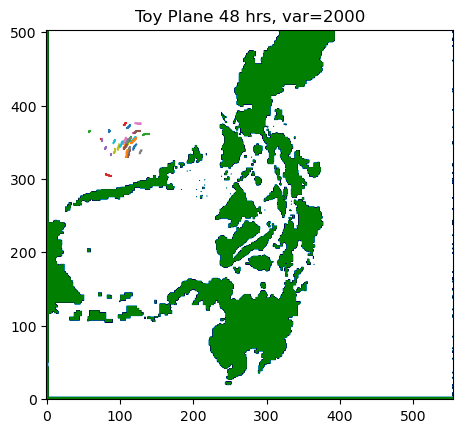
Three plots of simulation for Toy Plane crash location of x=350km, y=1050km

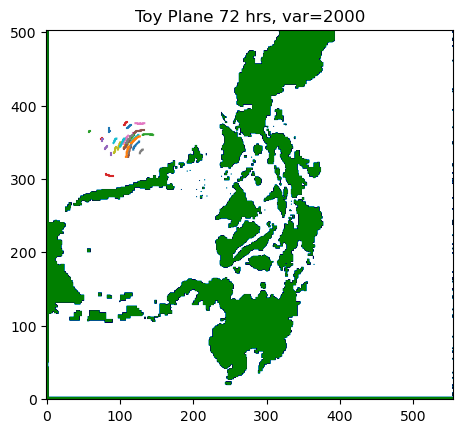


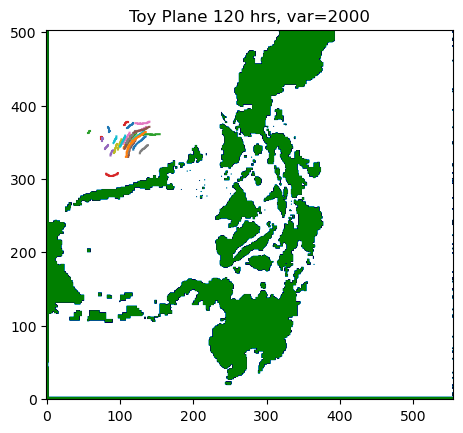




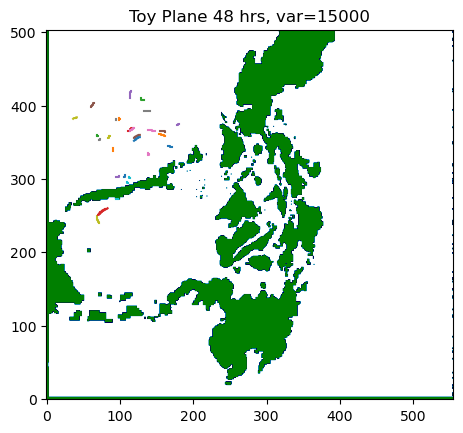
Three plots of simulation for Toy Plane crash location with small variance:

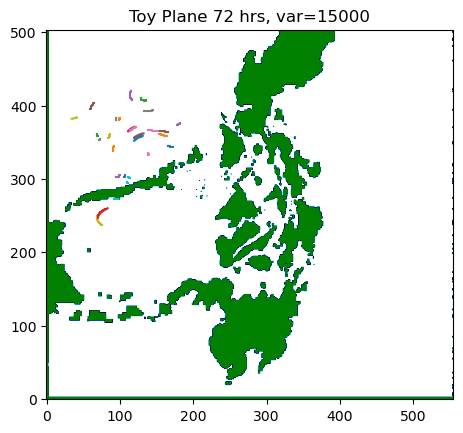


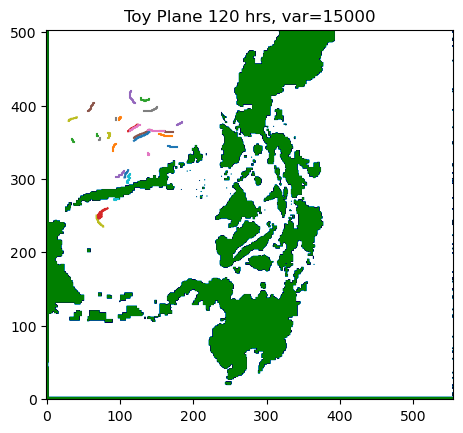




Three plots of simulation for Toy Plane crash location with large variance:







After reviewing the simulation results, it appears the toy plane could not have traveled very far in 72 hours. Recommend search area to be the upper left quadrant of map.

**Problem 4.a**. *Build Gaussian Model - Find Kernel Parameters*

* ***(1 point)****: States the choice of kernel function and provides a justification for this choice.*
* ***(1 point)****: Identifies the parameters of the kernel function.*
* ***(1 point)****: Explicitly states the search space for each kernel parameter.*
* ***(1 point)****: Explicitly states the number of folds () for the cross-validation.*
* ***(3 points)****: Provides the optimal kernel parameters from the search.*
* ***(3 points)****: Provides a plot of the computed cost/performance metric over the search space for the kernel parameters.*

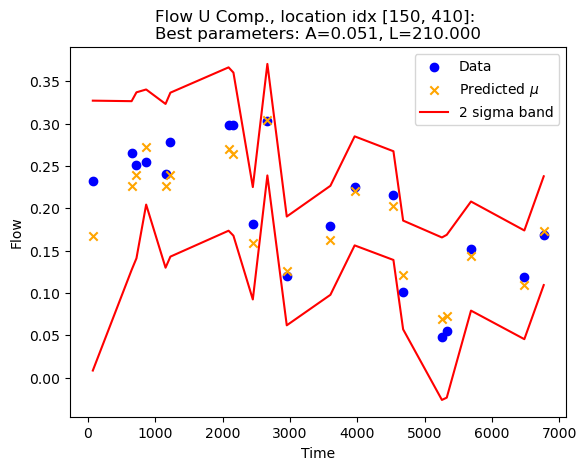
Kernel Function Choice: I chose the RBF Kernel because it is based on distance (time difference) and is exponentially decaying (correlation should drop after a short distance)

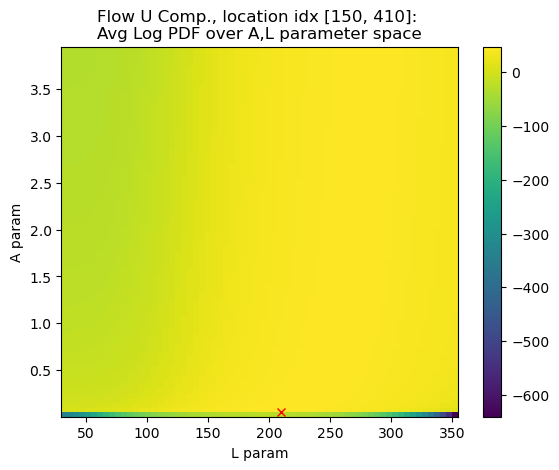
Kernel Parameters: The parameters to find are A and L, from kernel function:

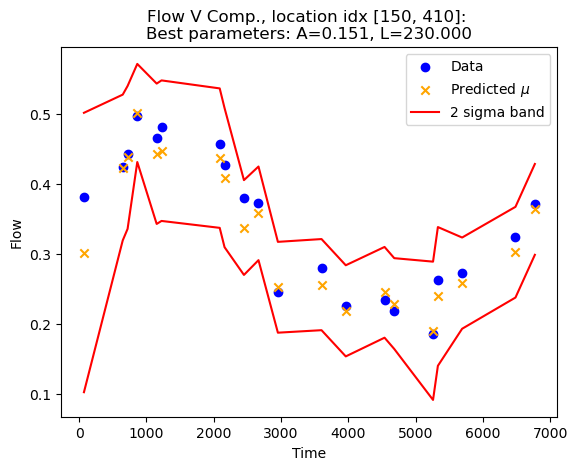
A\*exp( - ( ||z1-z2|| ^2 ) /( 2\* L^2 )

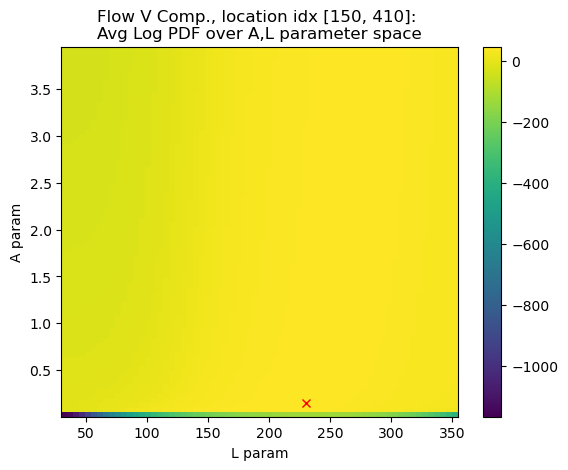
Search Space: range of A space is .001 to 4. range of L space is 30 to 360. Number of folds is 5.

The optimal Kernel Parameters A and L found are show for each U and V Flow Component, for one location. First plot is the conditional mean and observed data, Second plot is the log likelihood (cost metric) for the parameter search space. Marked in red x is the optimal A and L parameter.



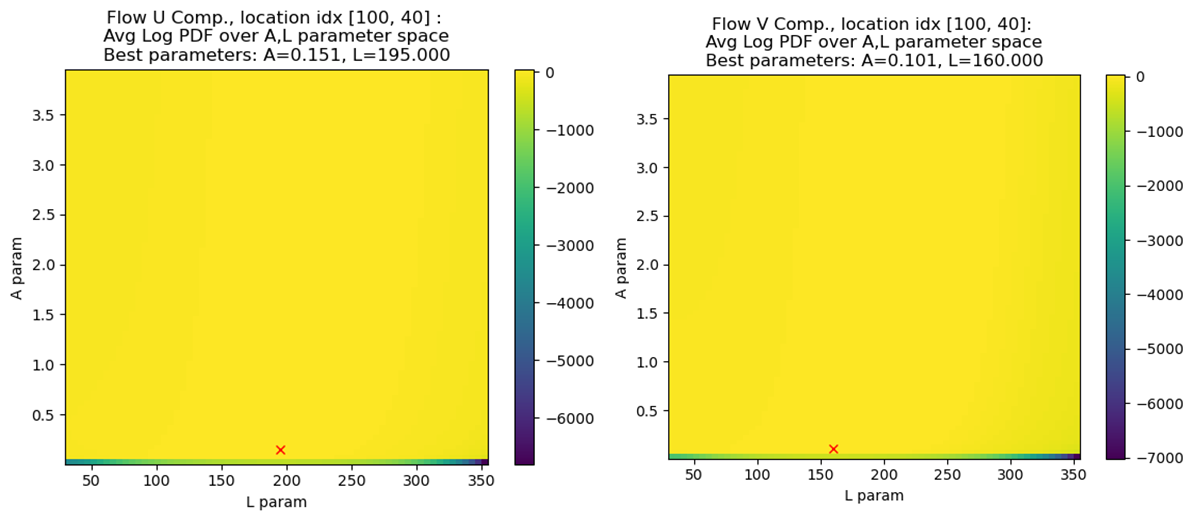


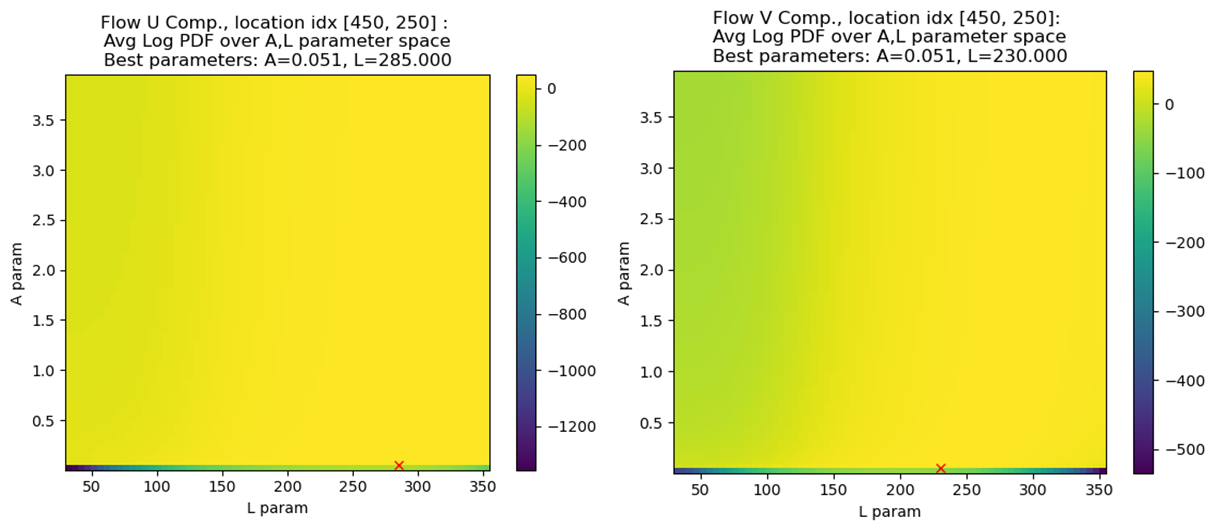


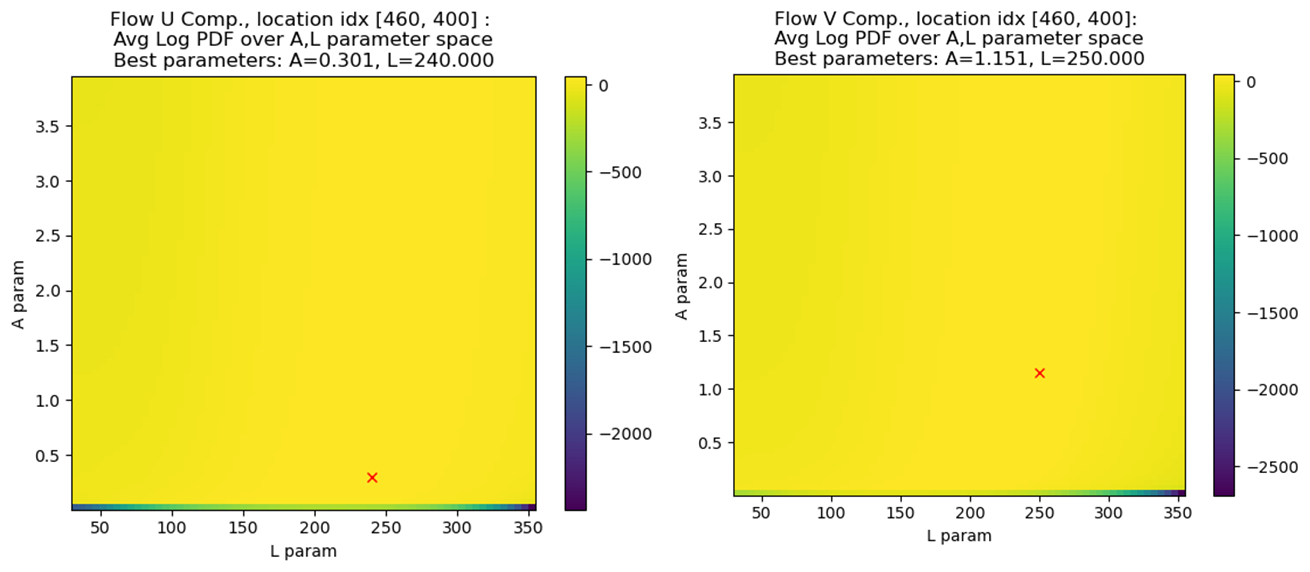


**Problem 4.b** *Provide best kernel parameters for 3 other locations*

* ***(3 points)****: Provides the optimal kernel values for three new location that are different from the location in Problem 4.a. (Plots do not need to be provided.)*
* ***(2 points)****: For each kernel parameter, states if a pattern was observed.*





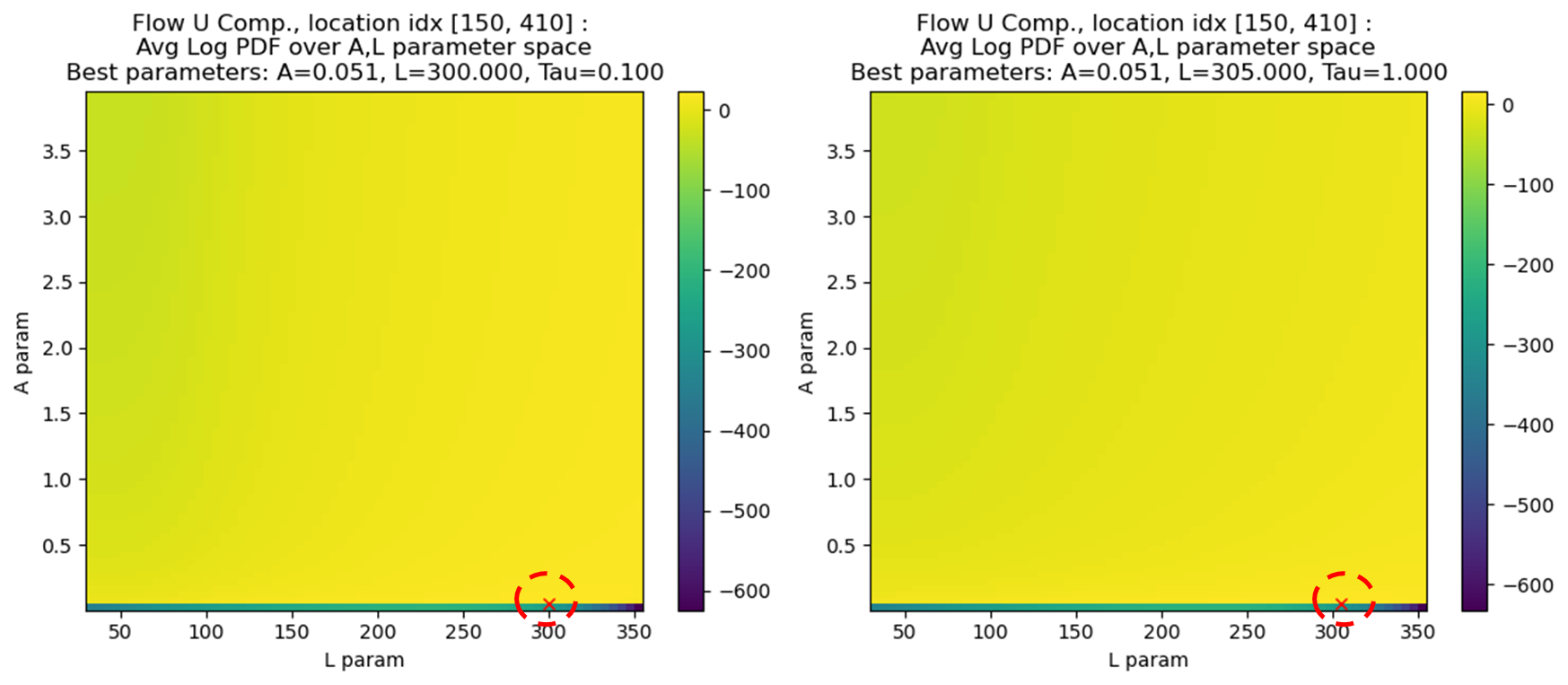


The three additional locations were chosen in different corners of the map. Yet the parameters A and L appear to remain roughly in the same space, A is mostly below 0.5, L is between 200 and 300.

**Problem 4.c:** *Tau Parameter on Kernel Function*

* ***(1 point)****: Provides the optimal kernel values for at least two new choices of .*
* ***(2 points)****: A plot showing the cost/optimization target is provided for the search space, for each choice of .*
* ***(2 points)****: Comments on whether these results differ from those found in Problem 4.a, and on whether results from the choices of in the problem differ from each other.*

Tried new values for Tau (0.1 and 1.0). The resulting best parameter A stayed the same, but L increased with increasing Tau.



**Problem 5** [15 pts] *Estimating unobserved flow data*.

* ***(2 points)****: Clearly states the choice of time-stamps at which to create predictions,* ***and*** *states why the choice was made.*
* ***(2 points)****: Clearly states the method by which the prior means were chosen.*
* ***(2 points)****: Provides a plot with a prediction for the* ***horizontal*** *velocity component at the chosen location.*
* ***(2 points)****: Provides a plot with a prediction for the* ***vertical*** *velocity component at the chosen location.*
* ***(3 points)****: Both plots have a labelled prediction for the mean for all of the time-stamps chosen.*
* ***(3 points)****: Both plots have a labelled band around the predicted mean for all of the time-stamps chosen.*
* ***(1 point)****: Both plots have the observations included.*

Timestamps: I chose to use a timestamp per day: where each 3rd day has an observed data, and the 2 days between are unobserved. This is more challenging than having only one unobserved data between each observed day.

Priors: I chose to use a prior mean of zero because in real world applications it is likely you will have limited data.

