MITx 6.419x Data Analysis: Statistical Modeling and Computation in Applications
Written report -- Homework 5
Student Name (Xing Zhang)

5/31/21

Module 5: Environmental Data and Gaussian Processes

Part I - Ocean Flow

Problem 2: Identifying long-range correlations

1. In this problem, we will try to identify areas in the Philippine Archipelago with long-range correlations. Your task is to identify two places on the map that are not immediately next to each other but still have some high correlation in their flows. Your response should be the map of the Archipelago with the two areas marked (e.g., circled). You claim that those two areas have correlated flows. Explain how you found that those two areas have correlated flows.

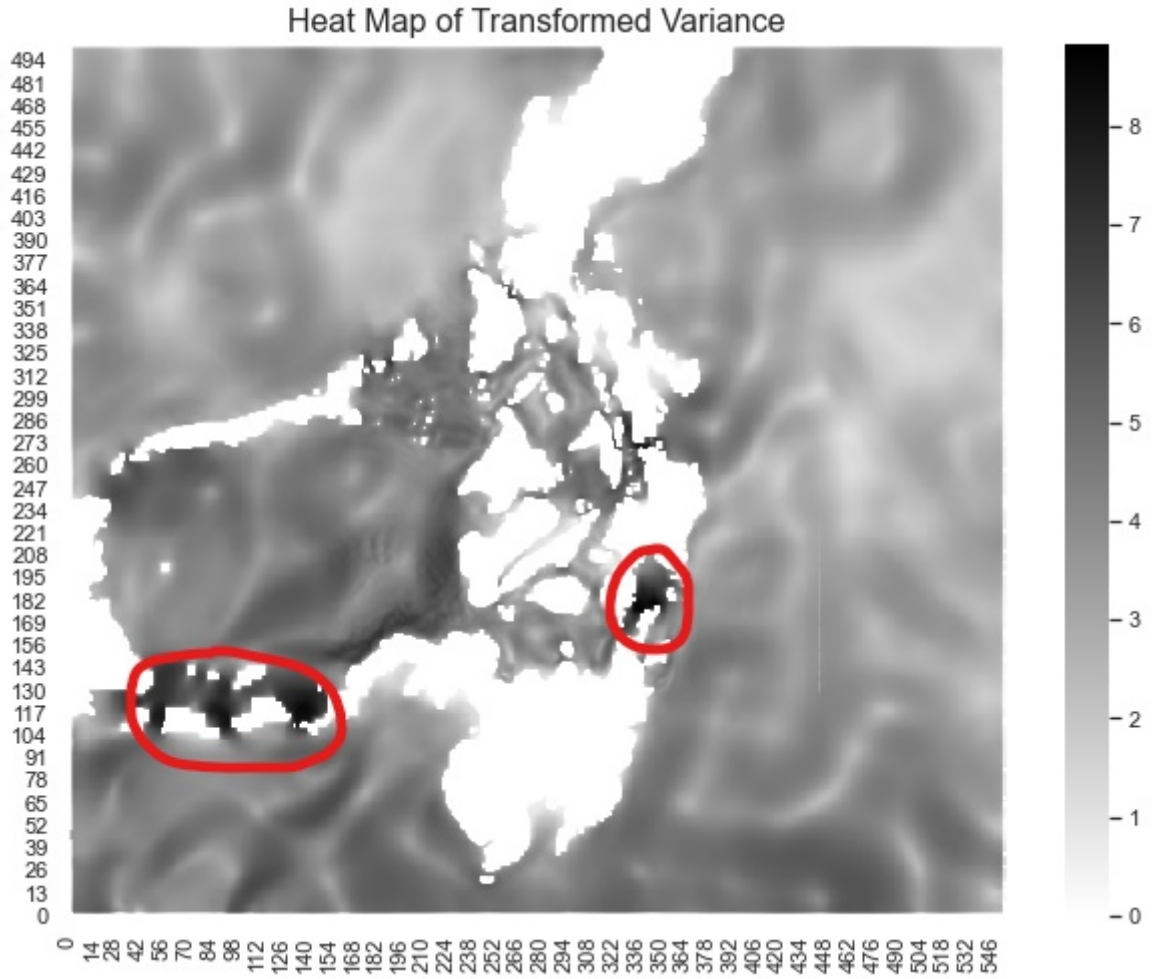
- A map with the two points with correlations marked.
- Provides an explanation of how the correlation was computed.
- Provides a convincing commentary on why the two marked locations could be correlated.

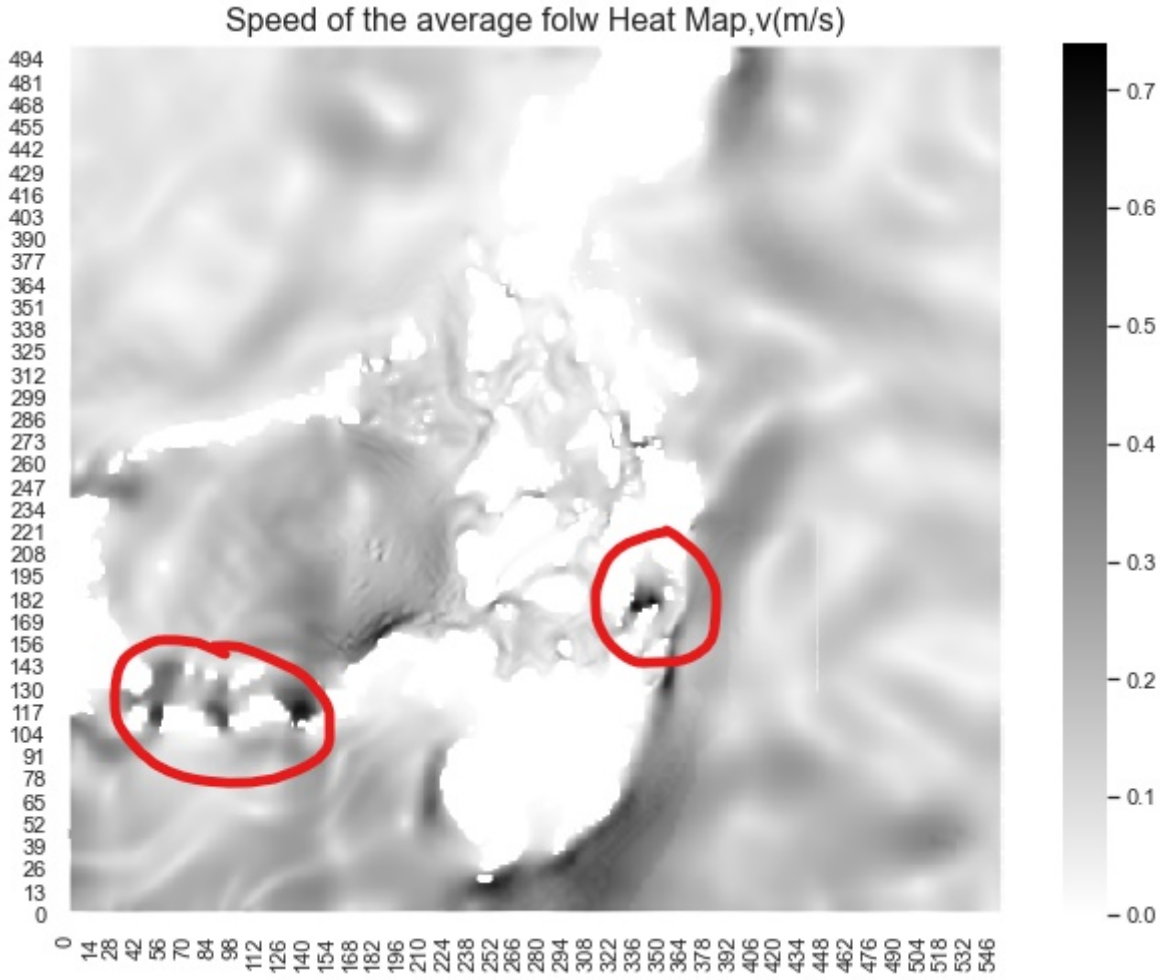
Solution:

Here, we will judge the correlation from two aspects: the total flow velocity value and the variance of the flow velocity change.

The total flow velocity $v = \sqrt{(V^2 + U^2)}$
The variance of the flow velocity change over the time can be calculated by `np.mean()`

The two points marked in the figure have highly correlate on flow velocity and variances. From the geographical location in the figure, it can be found that the two locations are in the strait. When the sea water passes through the narrow land channel, the pressure causes the flow velocity to increase. Due to the influence of tidal forces, the flow direction at this location also varies greatly, so the flow velocity variance It's also very big.


Heat Map of Transformed Variance
494 481 468 455 442 429 416 403 390 377 364 351 338 325 312 299 286 273 260 247 234 221 208 195 182 169 156 143 130 117 104 91 78 65 52 39 26 13 0
0 14 28 42 56 70 84 98 112 126 140 154 168 182 196 210 224 238 252 266 280 294 308 322 336 350 364 378 392 406 420 434 448 462 476 490 504 518 532 546
8 7 6 5 4 3 2 1 0


Speed of the average folw Heat Map,v(m/s)
494 481 468 455 442 429 416 403 390 377 364 351 338 325 312 299 286 273 260 247 234 221 208 195 182 169 156 143 130 117 104 91 78 65 52 39 26 13 0
0 14 28 42 56 70 84 98 112 126 140 154 168 182 196 210 224 238 252 266 280 294 308 322 336 350 364 378 392 406 420 434 448 462 476 490 504 518 532 546
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

Problem 3: Simulating particle movement in flows

In this problem, you are asked to build a simulator that can track a particle's movement on a time-varying flow.

a. We assume that the velocity of a particle in the ocean, with certain coordinates, will be determined by the corresponding water flow velocity at those coordinates. Implement a procedure to track the position and movement of multiple particles as caused by the time-varying flow given in the data set. Explain the procedure, and show that it works by providing examples and plots.

Draw particle locations uniformly at random across the entire map, do not worry if some of them are placed on land. Simulate the particle trajectories for 300 hours and provide a plot of the initial state, a plot of the final state, and two plots at intermediate states of the simulation. You may wish to draw colors at random for your particles in order to help distinguish them.

- Provides an explanation of the simulation algorithm, with equations for the evolution of the particle trajectory.
- Provides a plot of the initial state of the simulation.
- Provides two plots of intermediate states of the simulation.
- Provides a plot of the final state of the simulation.

Solution:

Explanation of the simulation algorithm:

> A flow refers to displacement in time, that is if $x(t)$ denotes the location of a particle at time t. Then, we define the flow vector as
> \begin{equation}
> \frac{d x(t)}{dt}= V(x(t),t).
> \end{equation}
> or similarly, their explicit Euler discretization as, for a small $\varepsilon > 0$, then
> \begin{equation}
> x(t+\varepsilon) = x(t) + \varepsilon V(x(t),t).
> \end{equation}
> Informally, if we know the position of a particle at time t, then, after some arbitrarily small time ε, that is, at time $t + \varepsilon$, we can compute the new position $x(t + \varepsilon)$ as $x(t + \varepsilon) = x(t) + \varepsilon V(x(t), t)$, assuming we know the flow information coded in the function $V(x(t), t)$.
> The next figure shows a simple flow system with four points, shown as white circles. Each point corresponds to a physical location also shown in kilometers. Attached to each point is a flow data point, which is shown as a blue arrow. Recall that flow data is given in the x and y direction. It is assumed that a particle moving in one of the zones or boxes acquires the velocity given by the flow data.

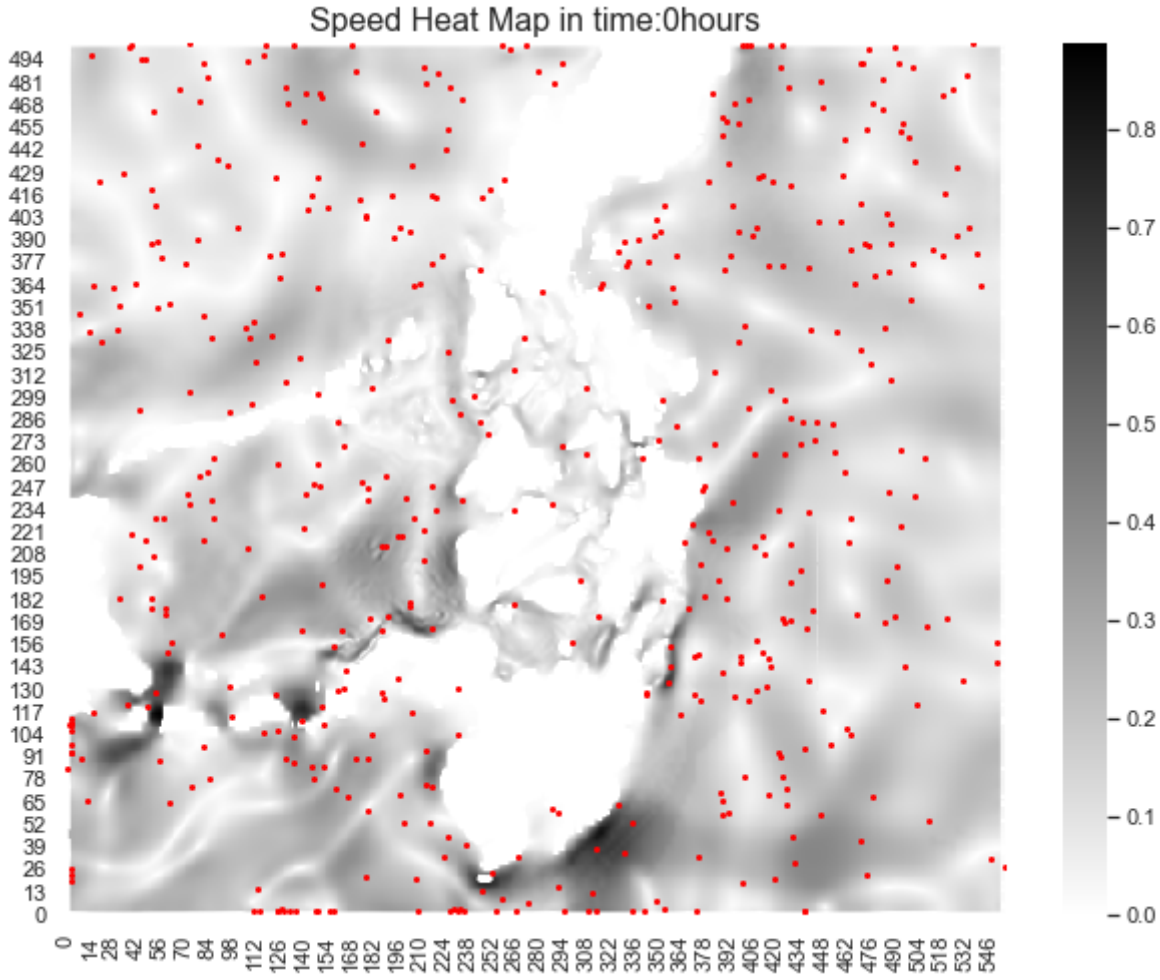
> As noted by the description above, in order to simulate the movement of the particle we need a couple of items:

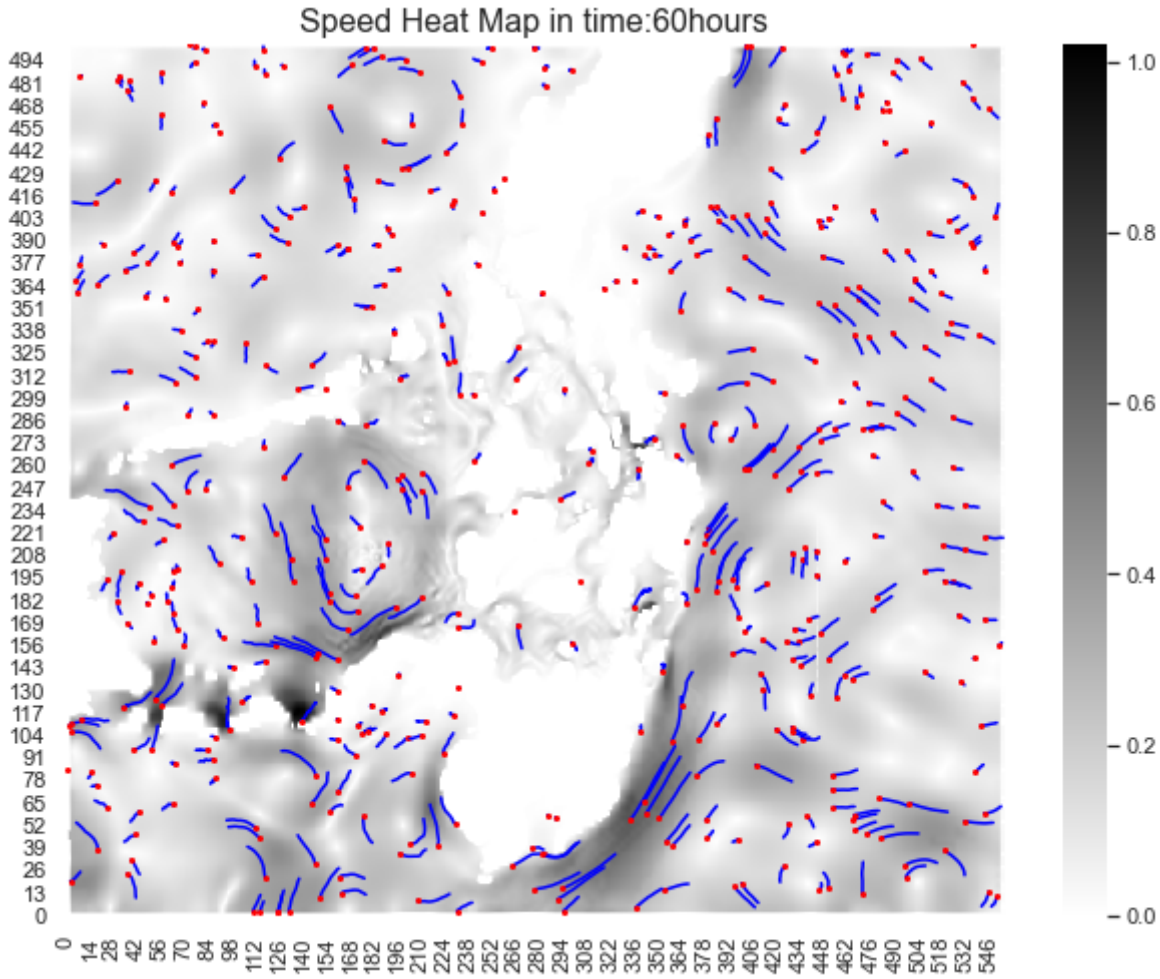
- A space $\mathcal{X}$ where we want out simulation to be run on. For example for a 2-dimensional space, we can consider a box with unitary lenght, i.e., $\mathcal{X} = [0, 1] \times [0, 1]$.
- An initial position $x(0) \in \mathcal{X}$. Of course we want our initial position to be inside the space we want to simulate the movement.
- A simulation time T. This is the maximum time we would like to run our simulation.
- A time step $T \geq \varepsilon > 0$. The unit ε is going to serve as a discretization of time, and is going to the the smallest unit of time we are going to consider in our simulation. In practice, this

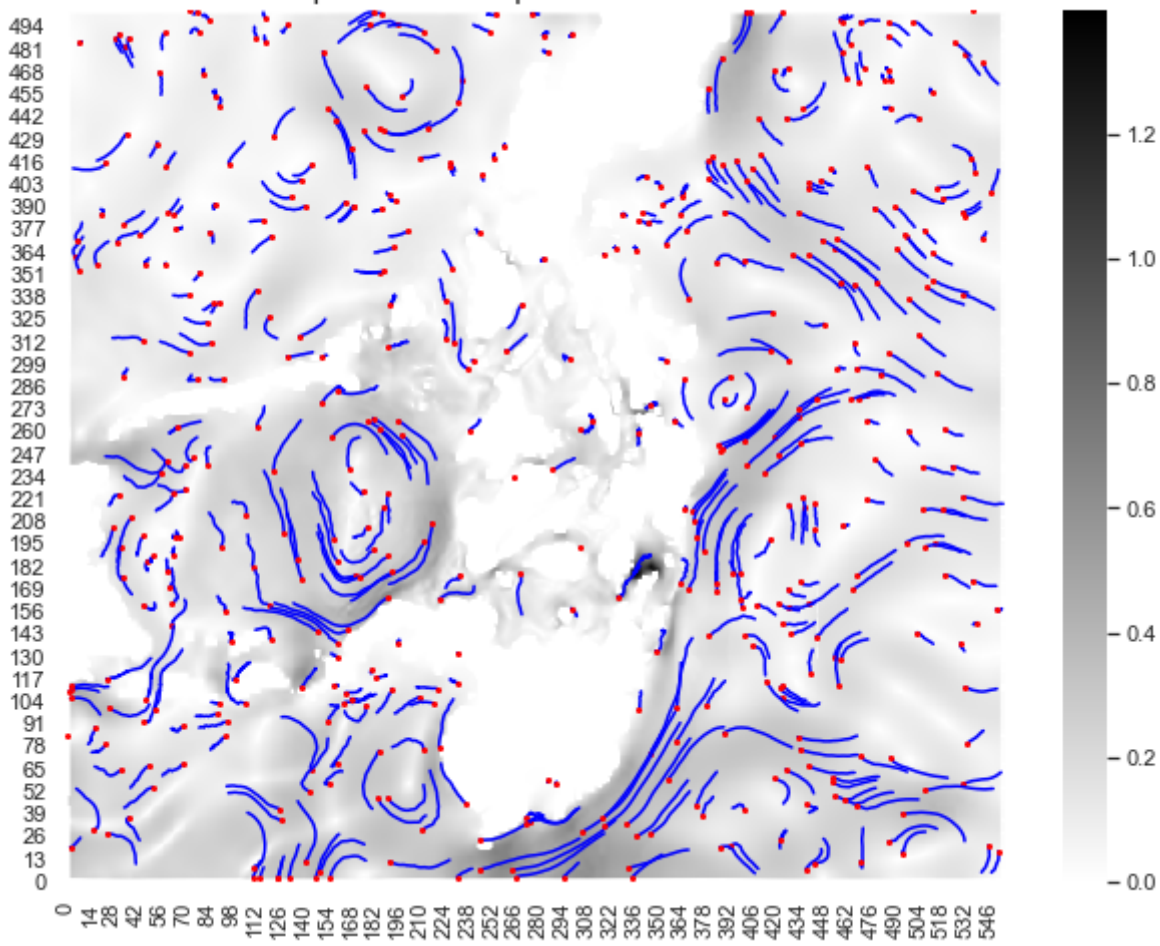
means our simulation should output a sequence of positions for times $\{0, \varepsilon, 2\varepsilon, 3\varepsilon, 4\varepsilon, \ldots, T\}$.

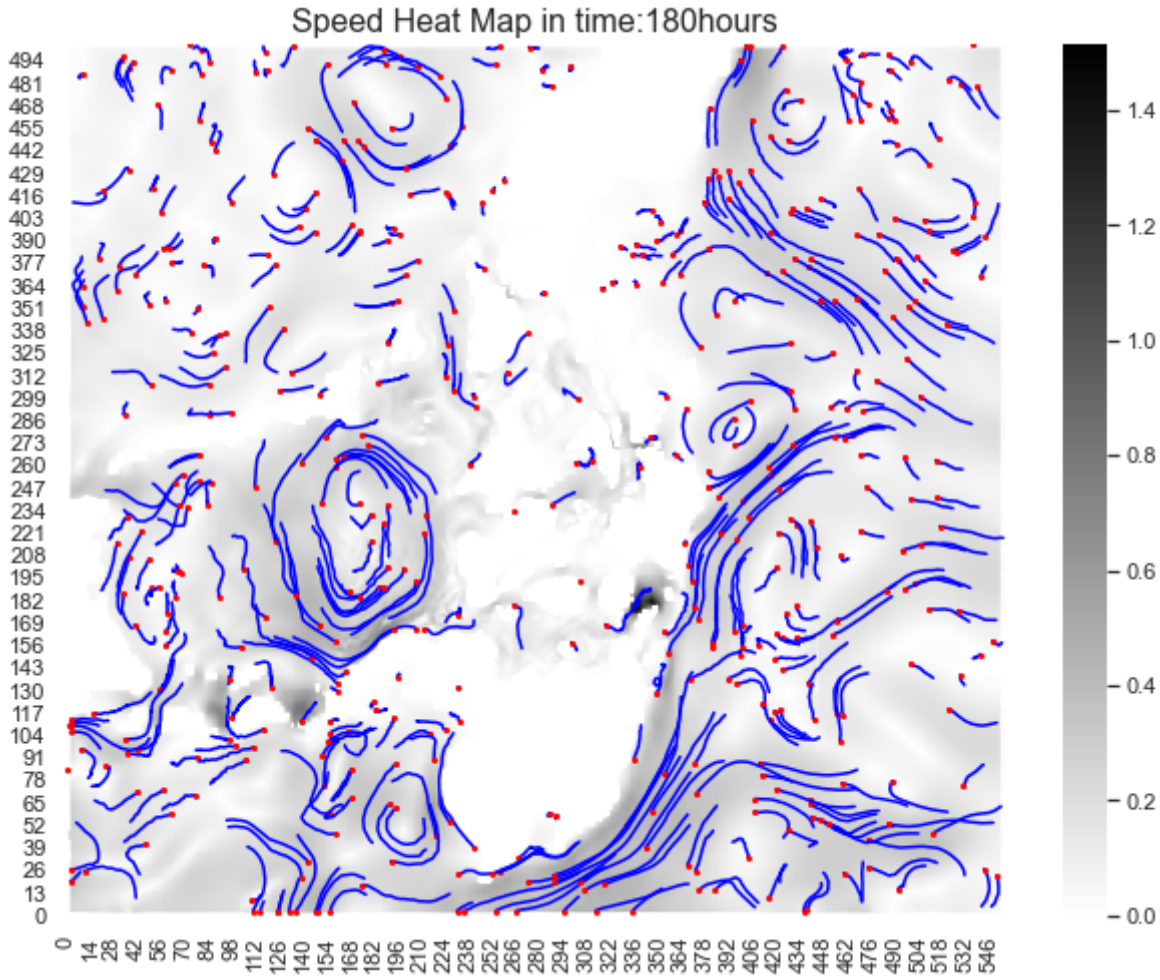
- The number ε also defines the number of iterations we are going to have in our simulation. For example, if $T = 10$, and $\varepsilon = 2$, then we will run a total of $N = 5$ iterations. One can start the oposite direction, by defining the number of iterations first and then compute the corresponding time step. For example, if I want to run $N = 300$ iterations, I will have to set \begin{equation} \varepsilon = \frac{T}{N} = \frac{10}{300} \approx 0.033.\end{equation}
- A flow function $V(x(t), t) : \mathcal{X} \times [0, T] \to \mathbb{R}^2$. This function has two input arguments, namely, a position $x(t)$ and a time t, and outputs a vector in $\mathbb{R}^2$ which indicates the velocity that a particle at location $x(t)$ would have in the x and y axis respectively. Note that we have explicitly allowed the flow function to depend on time. That is, flows might change depending not only on the location but of the time index.
- In general, we might not have access to a generic function of the form $V(x(t), t) : \mathcal{X} \times [0, T] \to \mathbb{R}^2$, specially if our flow comes from data that has been obtained at some particular set of locations. In this case, we might consider a set $\mathcal{G} = \{x_1, x_2, \ldots, x_m\}$ where $x_i \in \mathcal{X}$ for $i \in \{1, 2, \ldots, m\}$. These points represents the locations where flow information is available. For simulation purposes in this case, if at some point in time $t \in [0, T]$, the particle is at a location $x(t)$, instead of using the flow $V(x(t), t)$ which might not be available, we will use the surrogate flow $V(x_j, t)$ where $x_j = \operatorname{argmin}_{z \in \mathcal{G}} d(z, x(t))$ where $d(\cdot, \cdot)$ is some distance function, for example $d(x, y) = \|x - y\|^2$. This indicates that we will use the flow information at a point x_j in the set of available locations, that is closest to $x(t)$ in some predefined sense.

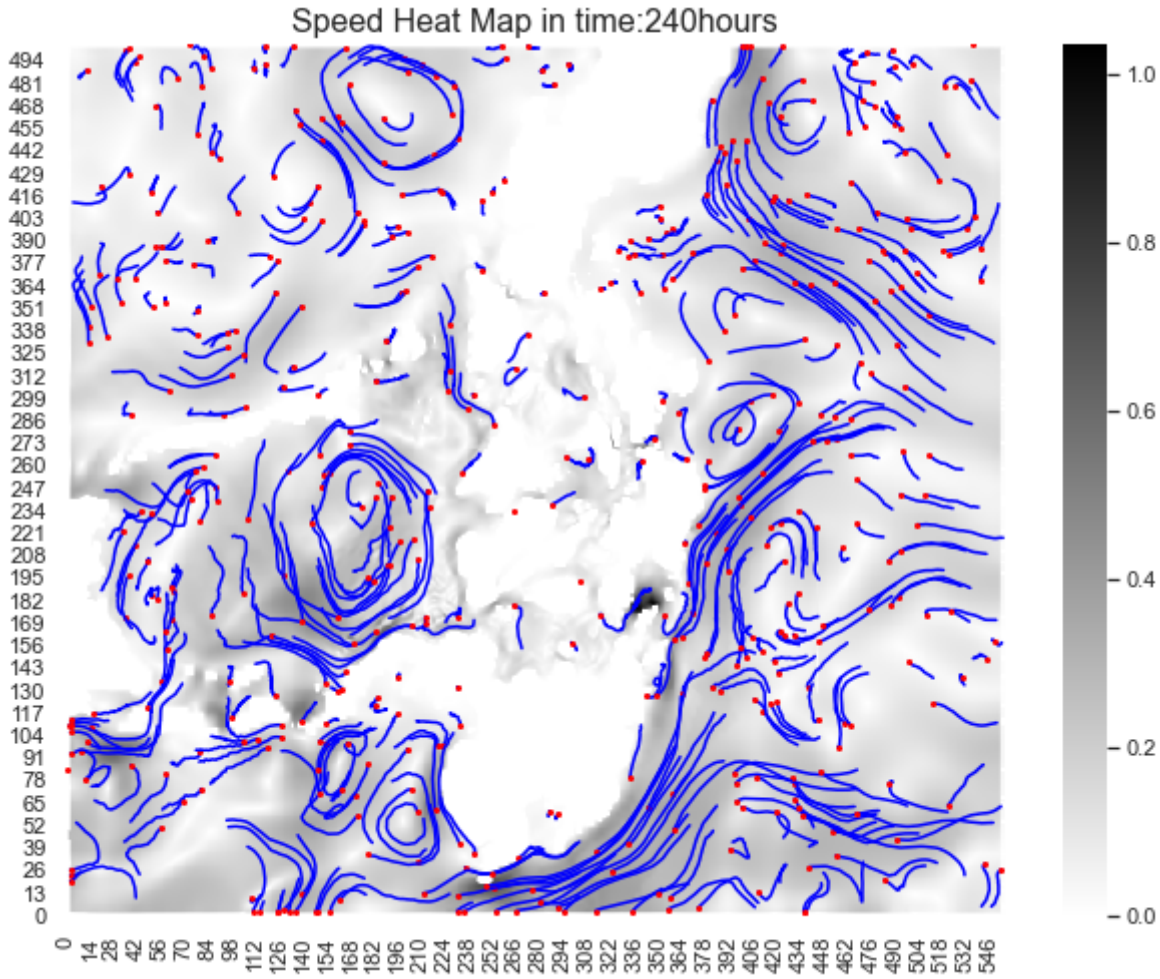
The trajectory is shown in the figure below, and the red mark is the end position at the corresponding time

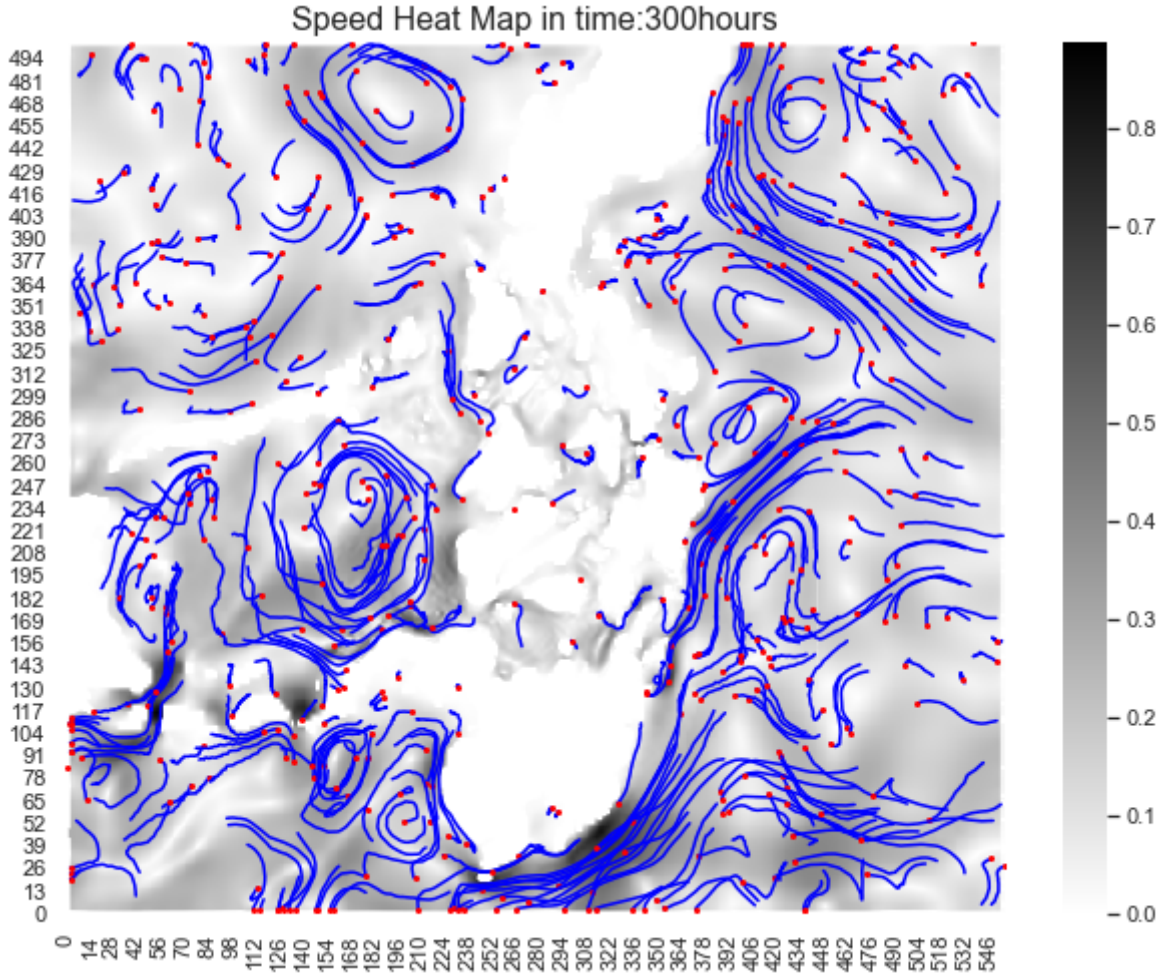



Speed Heat Map in time:60hours
494 481 468 455 442 429 416 403 390 377 364 351 338 325 312 299 286 273 260 247 234 221 208 195 182 169 156 143 130 117 104 91 78 65 52 39 26 13 0
0 14 28 42 56 70 84 98 112 126 140 154 168 182 196 210 224 238 252 266 280 294 308 322 336 350 364 378 392 406 420 434 448 462 476 490 504 518 532 546
1.0 0.8 0.6 0.4 0.2 0.0


Speed Heat Map in time:120hours
494 481 468 455 442 429 416 403 390 377 364 351 338 325 312 299 286 273 260 247 234 221 208 195 182 169 156 143 130 117 104 91 78 65 52 39 26 13 0
0 14 28 42 56 70 84 98 112 126 140 154 168 182 196 210 224 238 252 266 280 294 308 322 336 350 364 378 392 406 420 434 448 462 476 490 504 518 532 546
1.2 1.0 0.8 0.6 0.4 0.2 0.0


Speed Heat Map in time:180hours
494
481
468
455
442
429
416
403
390
377
364
351
338
325
312
299
286
273
260
247
234
221
208
195
182
169
156
143
130
117
104
91
78
65
52
39
26
13
0
0
14
28
42
56
70
84
98
112
126
140
154
168
182
196
210
224
238
252
266
280
294
308
322
336
350
364
378
392
406
420
434
448
462
476
490
504
518
532
546
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0


Speed Heat Map in time:240hours
494 481 468 455 442 429 416 403 390 377 364 351 338 325 312 299 286 273 260 247 234 221 208 195 182 169 156 143 130 117 104 91 78 65 52 39 26 13 0
0 14 28 42 56 70 84 98 112 126 140 154 168 182 196 210 224 238 252 266 280 294 308 322 336 350 364 378 392 406 420 434 448 462 476 490 504 518 532 546
1.0 0.8 0.6 0.4 0.2 0.0

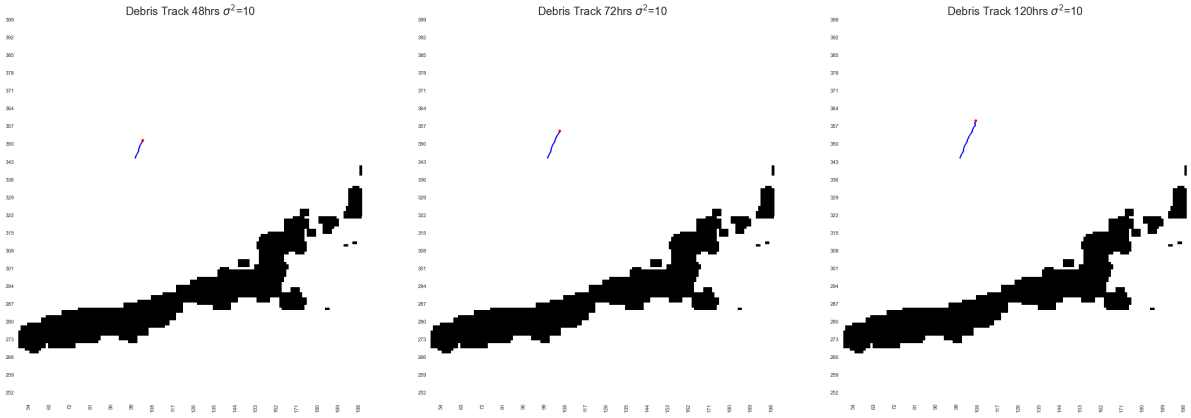

Speed Heat Map in time:300hours
494 481 468 455 442 429 416 403 390 377 364 351 338 325 312 299 286 273 260 247 234 221 208 195 182 169 156 143 130 117 104 91 78 65 52 39 26 13 0
0 14 28 42 56 70 84 98 112 126 140 154 168 182 196 210 224 238 252 266 280 294 308 322 336 350 364 378 392 406 420 434 448 462 476 490 504 518 532 546
0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

b. A (toy) plane has crashed in the Sulu Sea at $T = 0$. The exact location is unknown, but data suggests that the location of the crash follows a Gaussian distribution with mean $(100, 350)$ (namely $(300km, 1050km)$) with variance σ^2. The debris from the plane has been carried away by the ocean flow. You are about to lead a search expedition for the debris. Where would you expect the parts to be at 48hrs, 72hrs, 120hrs? Study the problem by varying the variance of the Gaussian distribution. Either pick a few variance samples or sweep through the variances if desired. (Hint: Sample particles and track their evolution.)

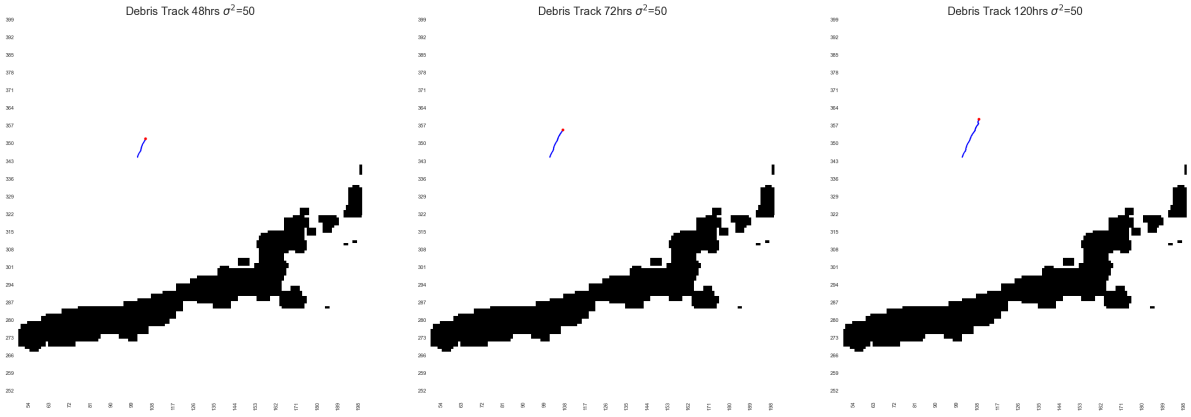
- Provides plots showing the state of the simulation at the times: T=48hrs, 72hrs, 120hrs. (Three plots required.)
- Two or more additional choices of the variances were tried, and three plots of the state of the simulation at the above three times are provided. (Six additional plots required.)
- Comments on where one should concentrate search activities based on the observed results.

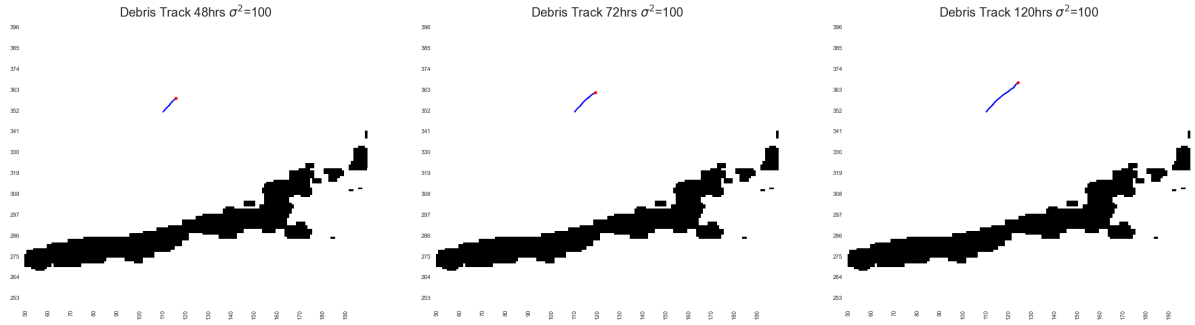
Solution:

(1)The red mark represents the end position of the toy plane drifting after a period of time.σ^2=10

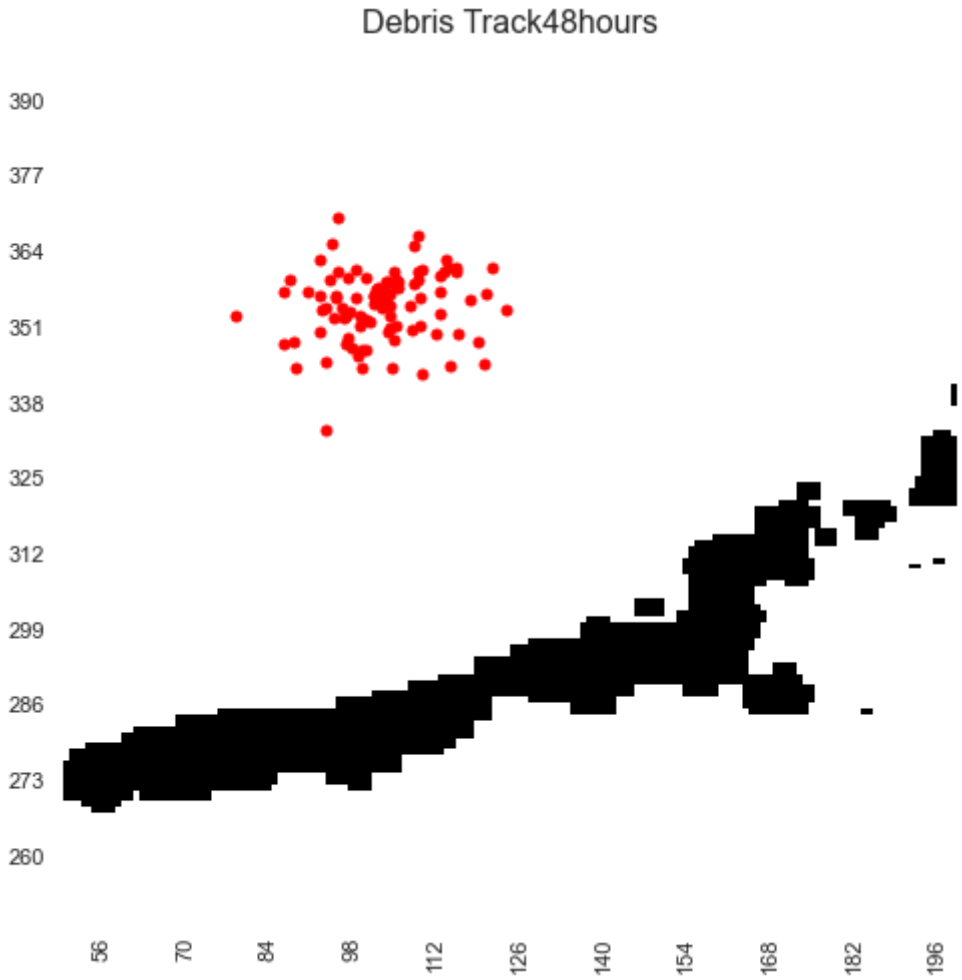


(2)The red mark represents the end position of the toy plane drifting after a period of time.σ^2 =50,100.

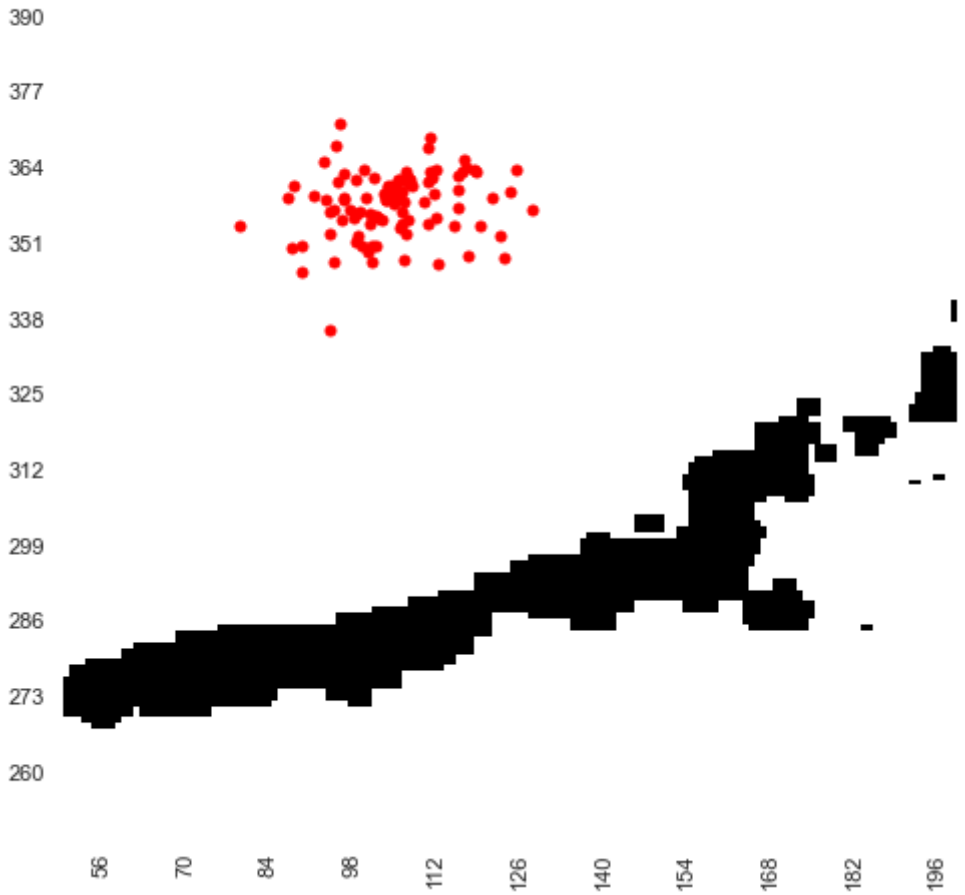




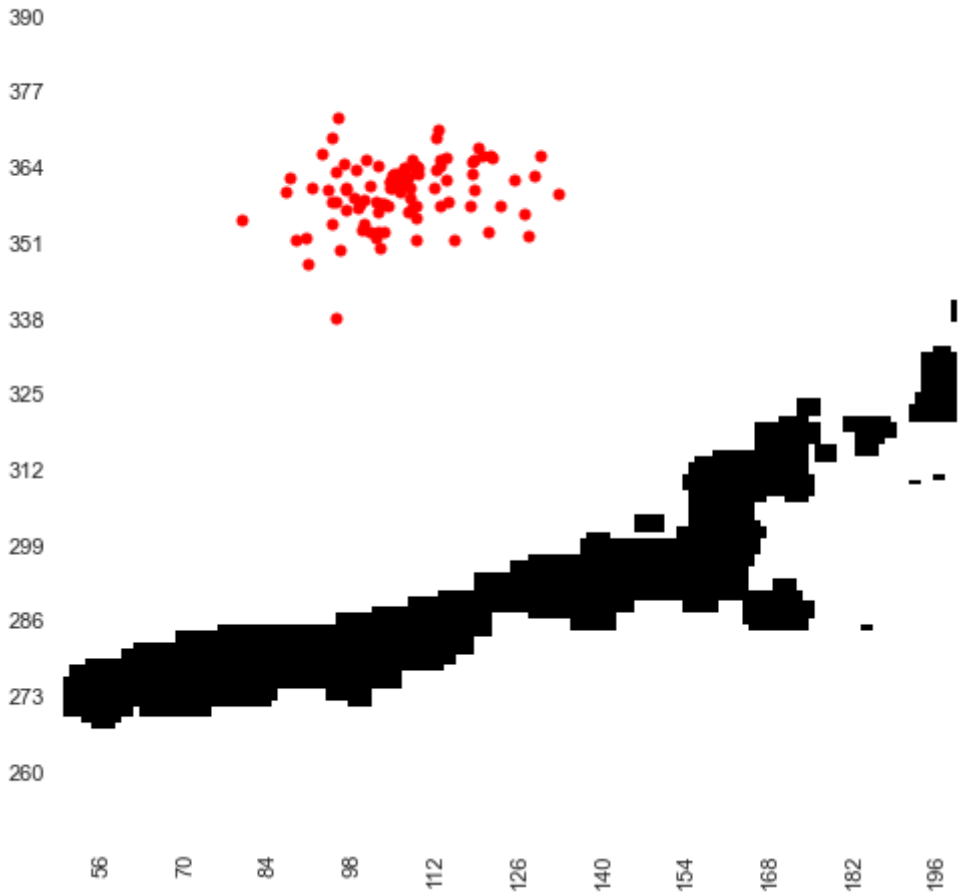
(3) Draw the search range of random falling locations with a variance value of 1-100. The red dots are the drifting end positions after a period of time. The toy plane should be found in the area enclosed by these red dots.

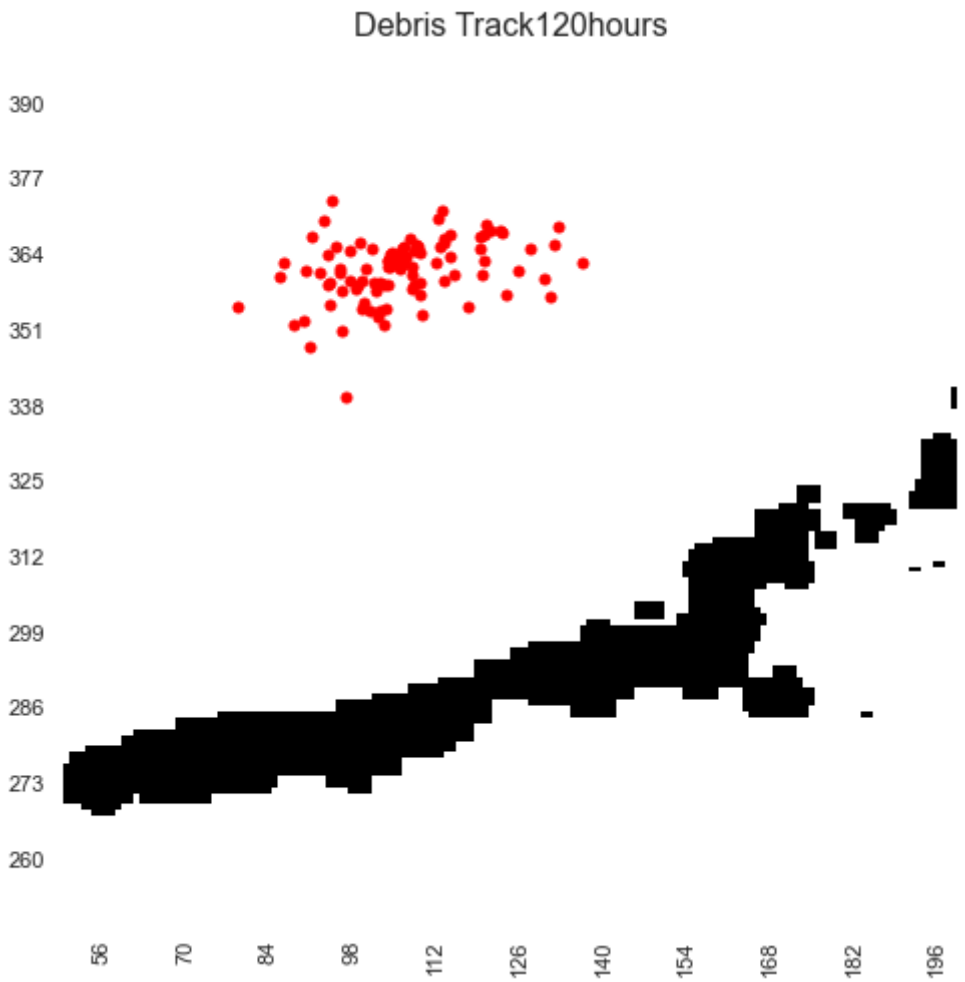


Debris Track72hours



Debris Track96hours





Part II - Estimating Flows with Gaussian Processes

Problem 4: Creating a Gaussian process model for the flow

In this problem, we will create a Gaussian process model for the flow given the information we have.

a. Pick a location of your liking from the map for which you are given flow data (ideally from a location on the ocean not in the land). Moreover, consider the two vectors containing the flow speed for each direction: you will end up with two vectors of dimension 100.

You are asked to find the parameters of the kernel function that best describes the data independently for each direction.

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

Solution:

Select the coordinate point at (x, y) = (320, 25)

(1) pick the kernel function:

$$k(z_1, z_2) = a \times \exp\left(-\frac{\|z_1 - z_2\|}{2l^2}\right).$$

The selected kernel function depends on a "distance" between the random variables at hand. In this case, each random variable represents the velocity. Of couse, this distance is in the time variable.

(2) For the selected kernel function, we can identify the set of parameters $\theta = (a, l)$.