Exercises

1	2	3	4	5	6	7	8	9	10

Surname, First name

_____

Modelling, Uncertainty and Data for Engineers (CEGM1000)
Resit 22/23 Q1

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Do not open the exam until given permission by the instructor!
(You can already write your name and student ID on this page)

The exam is 180 minutes. The table below gives an overview. On the following pages, some questions have a specific box for you to answer: anything written outside the boxes will not be graded. Note that we have provided a lot of space for answers. The answer space size is not an indicator of how long we expect your answers to be! (shorter is generally better). Scratch paper is available to use during the exam, but will not be collected or graded. You may use pen or pencil and a scientific (non-graphing) calculator. Separate formula sheet is provided.
Don't forget to write your student ID and fill in the bubbles on the top right of this page. Good luck!

No.	Question	Sub-Q	Type	Points
1	Coding	a	Open	10
		b, c, d,e	MC	8
2	Probability	a, b	MC	6
2		c	Open	2
3	Probability	a	MC	2
3		b	Open	4
4	Probability	a, b, c, d, e, f	Open	12
5	Mathematical Modelling	a, b, c, d	MC	8
6	Numerical Methods	a, b	Open	12
7	Sensing & Observation	a, b, c, d, e	Open	15
8	Sensing & Observation		Open	5
9	Simulation	a	MC	2
9		b	Open	3
10	Simulation	a, b	Open	6
		Total:		**95**

Part 1: Coding

The code below defines a ‘ParkingLot‘ Python class implementing the functioning of a parking lot in a simplistic way. The following description is provided from the documentation:

- ‘ParkingLot‘ objects are created by specifying the overall capacity of the parking lot, stored in ‘self.total_spots‘.
- ‘self.spots_available‘ is used to keep track of the current number of free spots; this value changes in time based on whether a vehicle leaves and enters the facility. Upon creation, ParkingLot objects are empty, i.e., there are no cars in the parking lot.
- ‘book_spot‘ and ‘free_spot‘ are used to either decrease or increase by one the count of spots available, e.g., whenever a vehicle exists or enters the parking lot. Checks are put in place with if statements to ensure that the ‘self.spots_available‘ counter does not go below zero or exceeds the parking lot capacity.
- ‘visualize‘ is used to print the current number of available spots, or that there are no spots available. This method is called once when the object is first created, and whenever the number of available spots changes.

10p **1a** There are **four different errors** in the implementation below. Localize these errors (you can use the line numbers as reference), define their type (e.g., syntax error, exception or logical error) and explain briefly how to fix them.

```
1  class ParkingLot:
2      """
3      A simple class implementing a parking lot.
4      """
5
6      def __init__(self, total_spots):
7          self.total_spots = total_spots
8          self.spots_available = 0
9          self.visualize()
10
11     def book_spot(self):
12         if self.spots_available >= 0:
13             self.spots_available -= 1
14             self.visualize()
15
16     def free_spot(self):
17         if self.spots_available < total_spots:
18             self.spots_available += 1
19             self.visualize()
20
21     def visualize(self):
22         if self.spots_available > 0:
23             print(f'There are [self.spots_available] spots available.')
24         else:
25             print('There are no spots available.')
```

1p **1b** With respect to the 'ParkingLot' code, which lines define the constructor of the class?

a Lines 1-9
b Lines 1-4
c Lines 6-9
d Lines 6-25

1p **1c** With respect to the 'ParkingLot' code, which line contains an attribute of the class?

a Line 6
b Line 7
c Line 19
d Line 21

3p **1d** Imagine that a correct implementation of the 'ParkingLot' class is available in the 'my_classes.py' file. Given the code below,

A
```
from my_classes import ParkingLot
parking_lot = ParkingLot()
```

B
```
import ParkingLot from my_classes
parking_lot = ParkingLot(20)
```

C
```
from my_classes import ParkingLot as ParkingLotClass
parking_lot = ParkingLotClass(20)
```

D
```
import my_classes
parking_lot = my_classes.ParkingLot(20)
```

which of the following statement is **FALSE**

(a) The import statement in (A) will correctly load the class
(b) (A) will fail to create an object due to an exception
(c) The outcome of running (C) and (D) is the same
(d) (C) will correctly load the class and create the object
(e) (B) will raise an exception

3p **1e** We want to add a check on the code defining the 'ParkingLot' class to limit the overall capacity of the parking facility to 200 parking slots. The code should stop and return an error when this occurs.

Which of these statements would implement the check correctly?
Regarding these results, mark **all** the options that are **TRUE**; consider that each wrong answer will result in negative points, but the lowest score for this sub-question is 0 (we will not subtract points from the rest of the exam):

A
```
assert total_spots <= 200, "Error: The parking lot is too large"
```

B
```
if total_spots > 200:
    print('Error: The parking lot is too large')
```

C
```
if total_spots > 200:
    raise ValueError('Error: The parking lot is too large')
```

D
```
if total_spots > 200:
    raise Exception('Error: The parking lot is too large')
```

☐ A ☐ B ☐ C ☐ D

Part 2: Probability

As an engineer in a consultancy firm, you need to perform the Extreme Value Analysis on the hourly wave height measurements from a buoy in the middle of the North Sea. You have applied both sampling methods in order to find a probability distribution to model the maximum wave height observed per year: Block Maxima and Peak Over Threshold. Recall that the objective is to select statistically independent and identically distributed samples; the Block Maxima method uses fixed increments of time and the Peak Over Threshold methods filter maxima using threshold values and a minimum time between peaks. In the following table, the outputs of the *describe* function (from *pandas*) applied on the sampled maxima are shown.

	Block Maxima	Peak Over Threshold
Count	7	35
Mean	5.81	5.09
Std	1.38	0.87
Min	4.63	4.25
25%	4.96	4.50
50%	5.14	4.89
75%	6.31	5.34
Max	8.39	8.39

4p **2a** Using the summary provided, which statement is the most valid?

(a) Block Maxima is the best method because it is a simple and fast method which ensures independent extreme observations.

(b) Block Maxima is the best method because wave height data is probably Gaussian distributed, so Generalized Extreme Value distribution would be a good fit for its extremes.

(c) Peak Over Threshold is the best method because it is a simple and fast method which ensures independent extreme observations.

(d) Peak Over Threshold is the best method because it maximizes the number of extremes sampled from our data and, in this case, our time series seems to be short.

Assume that a colleague did the sampling using the Peak Over Threshold method and asks your advice on the result, as shown in the plot below. The plot represents the sampled extremes for 3 weeks using different values for the threshold and the declustering time.

2p **2b** Based on such plot, which pair of values would you choose?

(a) Threshold = 3m and declustering time = 24h.

(b) Threshold = 4.25m and declustering time = 48h.

2p **2c** Explain why.

Part 3: Probability

For the same problem as in Part 2, assume that you now have the extreme observations. You want to fit a parametric distribution to them, and you can do it using different methods.

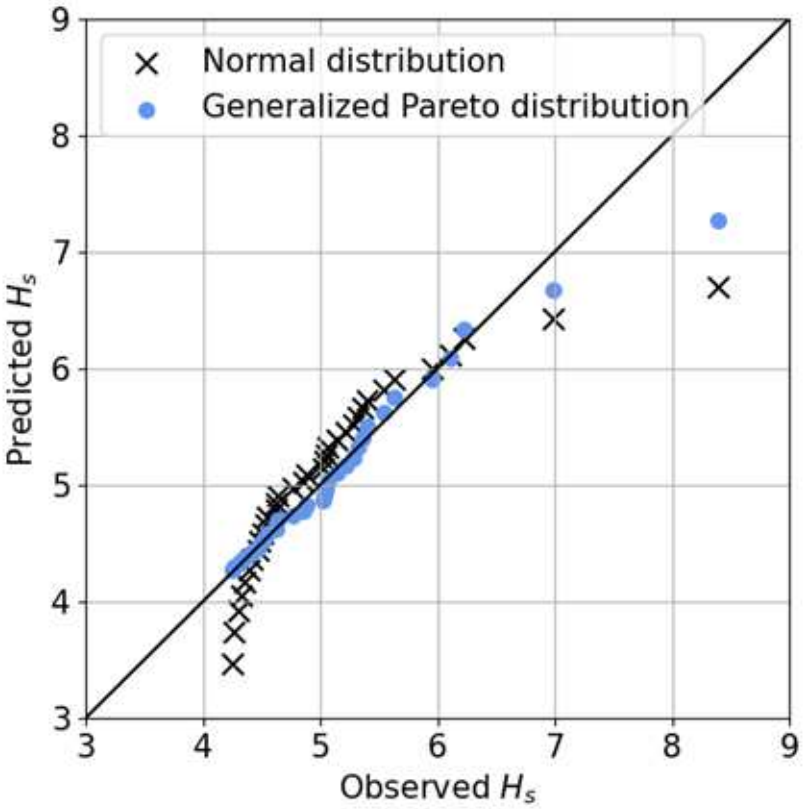
2p **3a** Which of the following methods cannot be applied to this end?

(a) Maximum Loglikelihood Estimator

(b) Inverse Modelling

(c) Using moments of data to match moments of the parametric distribution

Since you are not sure about the distribution that would be a good model for your data, you fit a Generalized Pareto distribution and a Normal distribution. After fitting them, you assess their goodness of fit using the Q-Q plot and the Kolmogorov-Smirnov test.



Kolmogorov-Smirnov test:

- Normal: p-value=0.294
- Generalized Pareto: p-value=0.892

4p **3b** Based on the results above, which distribution fits best your data? How have you determined it? Explain your answer.

Part 4: Probability

For the same problem as in Part 2 and 3, assume you choose the Gumbel distribution (see formula sheet).

2p **4a** Compute the distribution parameters using the information from Peak Over Threshold analysis. Round to two decimals.

In case you did not get an answer to the previous question, assume you obtained $\alpha = 1$ and $\beta = 2$.

2p **4b** Calculate $P(X > 5)$ using Gumbel distribution.

↳

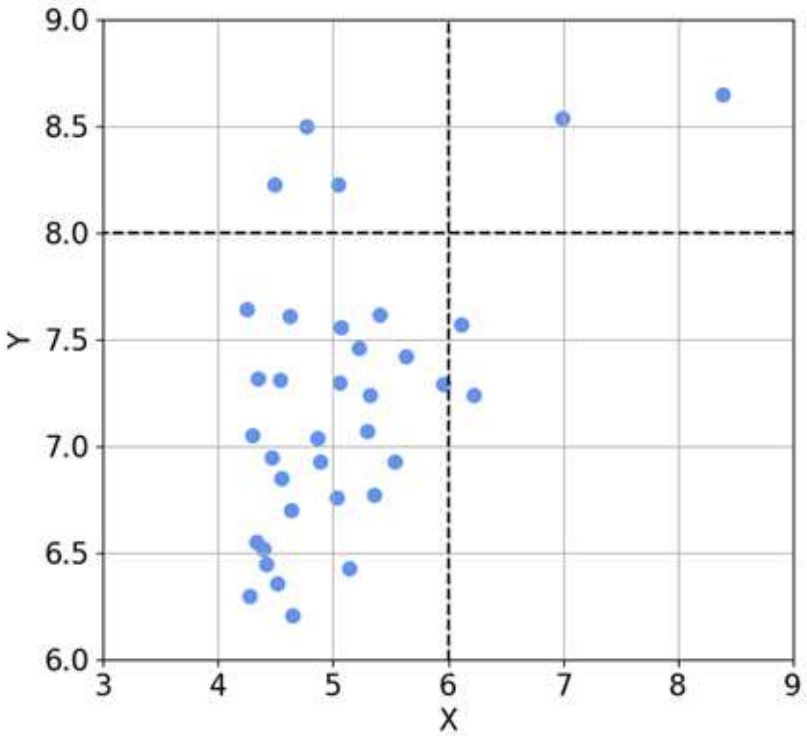
2p **4c** Calculate the value of the wave height (X) whose exceedance probability is 0.05.

You are also interested in the wave periods (Y) for your design. Assume that they follow an exponential distribution with $\lambda = 0.05$.

2p **4d** Assume that wave heights (X) and wave periods (Y) are independent, compute the probability of $P(X < 5, Y < 7)$.

2p **4e** Assume that wave heights (X) and wave periods (Y) are independent, compute the probability of $P(X < 5|Y < 7)$.

2p **4f** However, you know wave heights and periods are not independent. Given the plot below, calculate the empirical value of $P(X > 6, Y > 8)$. Note that there are 35 observations in the plot.



Part 5: Numerical Modelling: Validation and Verification

2p **5a** When can you actually say that a numerical model is verified?

(a) When the mathematical part of the model is correct, and the code is free of errors.
(b) When the model matches observed experimental data (or simulated data from validated models).
(c) When the numerical model matches analytical results obtained for simpler configurations of the system.

2p **5b** By definition, a sensitivity analysis implies:

(a) The validation of a model due to uncertainties of its parameters.
(b) Assessing the influence of certain parameters to the investigated response.
(c) Identifying code errors.
(d) Matching numerical modelling results to experimental results.

2p **5c** A sensitivity analysis is **not** useful for:

(a) Dimension reduction of the model.
(b) Model validation.
(c) Dividing and allocating the uncertainty of the output in our model to the different sources of variability of the input parameters.
(d) Implementation of a model

2p **5d** Which of the following is generally true? A model that has this form: $\dot{x} + 3x = 0$ and that has been derived from Newton's second law can be classified as

(a) A dynamic model, since it contains a first order derivative and depends on time.
(b) A static model, since it contains a first order derivative.
(c) A static model, since it doesn't contain a second order derivative.
(d) A dynamic model, since it has been derived from Newton's second law.

Part 6: Numerical Methods

6p **6a** Using Taylor expansion, derive the backward Euler approximation for the first derivative. Show the truncation error introduced by the approximation.

6p **6b** Derive the discrete form of the following ODE using the backward Euler approximation and calculate first 5 time steps of the solution using $dt = 0.2$:

$y' = y + t\cos t$
$y(0) = 1$

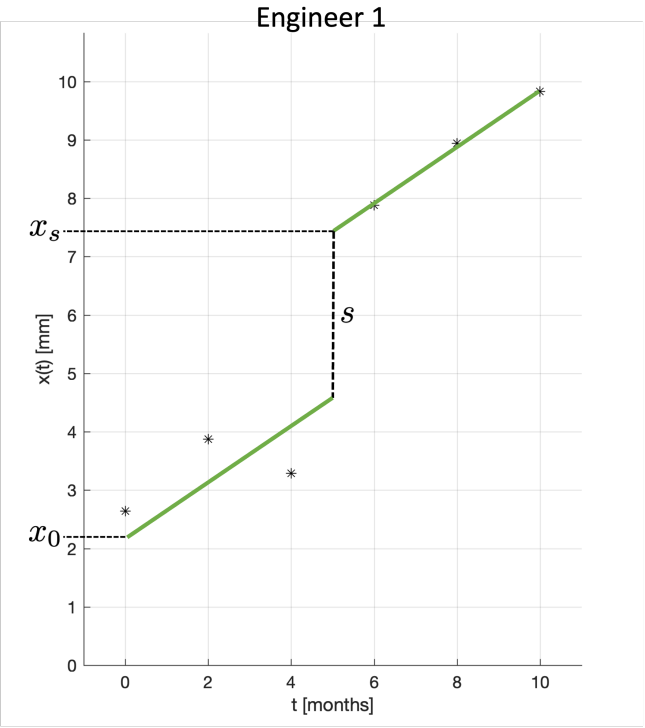
Part 7: Sensing and Observation

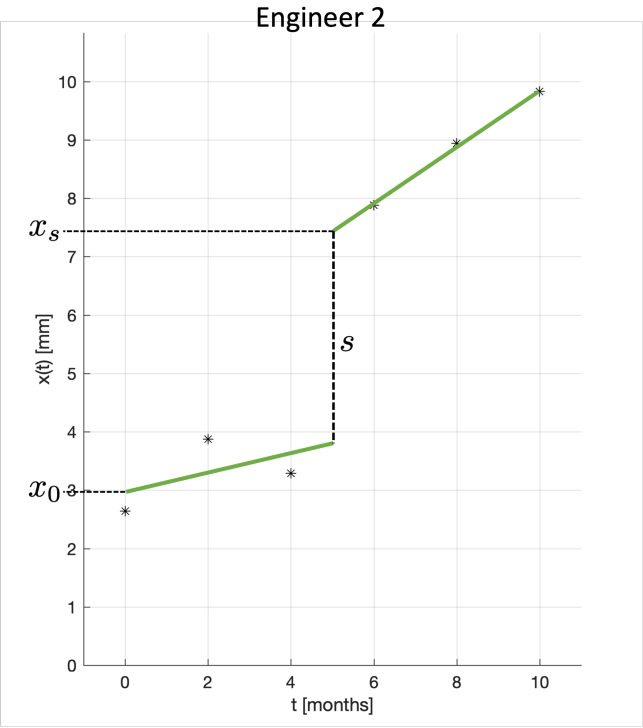
The distance x between a fixed benchmark and a moving benchmark on a landslide is measured at times $t = 0, 2, 4, 6, 8, 10$ months. The observations are shown in the figure. The precision of the first 3 observations is 0.5 mm, the precision of the last 3 observations is 0.2 mm. All observables are independent and normally distributed.

It is assumed that normally the distance is changing at a constant rate. It is known, however, that at $t = 5$ months there was a sudden slip of the landslide, causing an additional change in distance at that time.

Two engineers are asked to estimate the parameters that characterize the motion of the landslide.

- Engineer 1 uses as nominal model (null hypothesis) that before and after the slip, the rate of change was the same. The 3 unknown parameters she is estimating are: the distance x_0 [mm] at $t = 0$, the constant rate of distance change v_0 [mm/month], and the distance change s during the sudden slip event. See figure.
- Engineer 2 uses as nominal model (null hypothesis) that before and after the slip, the rate of change was different. This means that there will be 4 unknown parameters to estimate.





4p **7a** Give the functional model (in the form $\mathbb{E}(Y) = \mathrm{A} \cdot \mathrm{x}$) and stochastic model corresponding to the null hypothesis of Engineer 1.

4p **7b** Give the functional model corresponding corresponding to the null hypothesis of Engineer 2 in the form $\mathbb{E}(Y) = \mathrm{A} \cdot \mathrm{x}$. Explain (describe) what the 4 unknown parameters are.
Hint: there are multiple options, you only need to give one option. For example you can choose either x_s or s as one of the unknown parameters.

↳

2p **7c** Which test do the engineers need to apply to check the validity of their model?

2p **7d** What are the corresponding critical values used by the engineers with $\alpha = 0.05$?

3p **7e** Do you generally expect the test statistic of Engineer 1 to be larger or smaller than the test statistic of Engineer 2? Explain.

Part 8: Sensing and Observation

A company is producing instruments for measuring the CO2 concentration [ppm] in air. Their top range model has a precision of 2 ppm (for a single measurement). Scientists need to know the CO2 concentration with a 98% confidence interval of ± 0.75 ppm. Therefore they use repeated measurements to estimate the concentration.

5p **8** The company would like to improve their instrument, such that only 5 repeated measurements will be needed to meet the requirement. What is the required precision of a single measurement in that case?

Part 9: Simulation

Monte Carlo Simulations are a broad class of computational algorithms that rely on repeated random sampling to obtain a set of numerical results, and ultimately to quantify the uncertainty in the output.

2p **9a** State-of-the art algorithms for generating random samples from a distribution are based on the optimized pseudo-number generation of identically distributed samples. If the input variable has a Gaussian distribution, which is the most appropriate sampling method to be used?

(a) Inverse Transform

(b) Box-Mueller

(c) Standard Accept/Reject method

3p **9b** Why are the other two methods not appropriate? State one reason for each.

Part 10: Simulation

The second step of the Box-Mueller algorithm requires to compute

$r = \sqrt{(-2ln[1-u_1])}$
$\theta = 2\pi u_2$

with u being samples generated from the random variables:
$U_1 \sim U(0,1)$
$U_2 \sim U(0,1)$

r is often written in the code as
r = np.sqrt(-2 * np.log(U1))

3p **10a** Explain why it is correct to use the $log(u_1)$ instead of $log(1-u_1)$.

3p **10b** You have generated enough samples from $N(\mu, \sigma^2)$, so that your output mean is now converged to a value. Make two sketches of the empirical PDFs (probability density functions) for samples with a very small variance, and the other with a large variance.