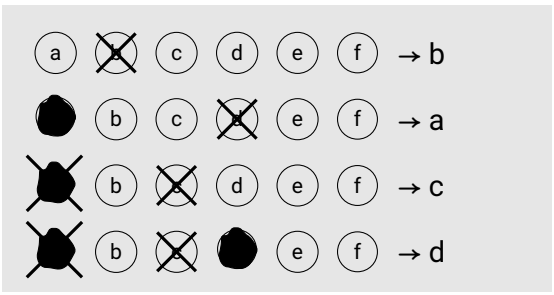
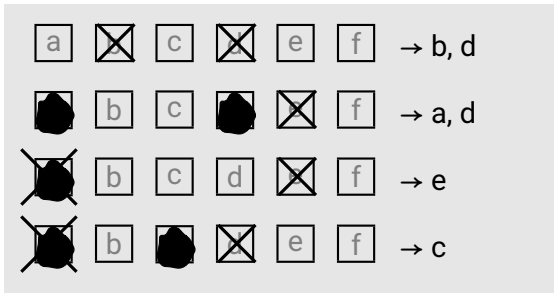
Exercises

1	2	3	4	5	6	7	8	9

Surname, First name

Modelling, Uncertainty and Data for Engineers EXAM (CEGM1000)
Exam Q2





Answer multiple-choice questions as shown in the example. Circular checkboxes have only one correct answer. Square checkboxes may have multiple correct answers.

===== DO NOT OPEN THIS EXAM OR TURN IT OVER UNTIL INSTRUCTED TO DO SO =====

Before you start the exam, a few remarks:

- Write down your **first and last name** in the field on the **top left** corner of this page.
- Write **student number** on the **top right** corner as the **number and filled in circle**.
- **New for this exam:** several questions ask require an answer on a figure. In each case, two identical images are provided in case you must correct your work. Cross over the image you do NOT want to submit; if it is not clear which image you want to submit, we will grade the bottom one.
- Place your student ID card face up on your desk.
- The duration of the exam is 3 hours (if entitled to an extension place paper on desk, face up).
- You are only allowed a pen/pencil and approved calculator; nothing else is allowed.
- Note in the multiple choice answer examples above that filled in circles/squares take precedence over an "X." **It is your responsibility to make sure that there is no ambiguity in your final answer**s (for example, add an extra note "final answer = x, y" etc). If in doubt, ask an instructor prior to submitting your exam. Examples are provided on the back page of this exam.
- Write all answers in the space provided (scrap paper is not graded), and do not remove the staple (nietje). In case you must erase, ask an invigilator for a white sticker to cover your incorrect answer. If you use any extra space to write an answer, indicate so in the original answer field.
- A summary of all questions is provided on the back page.

Good luck!

This page is left blank intentionally

Programming

A colleague sends you a piece of code in the file class.py, with instructions to import it using the following line of code:

```
from class import Class
```

3p **1a** Which line below would be the proper way to instantiate a class contained in class.py ?

(a) C = class

(b) class

(c) C = Class()

(d) Class()

(e) Instantiate(Class)

Consider the code snippet below, then answer the following **2** questions. Note that the file data.csv contains the date and daily measurements of an unknown property over a total period of 4.5 years.

```
D = pd.read_csv('data.csv', delimiter = ',', parse_dates = True)

indices = D.groupby(pd.DatetimeIndex(D['Date']).year)['olddata'].idxmax()

newdata = D.loc[indices]

print(newdata.shape)
```

3p **1b** The output from running this cell will **most likely** be:

(a) (5,)

(b) (5, 1)

(c) (5, 2)

(d) (54,)

(e) (54, 1)

(f) (54, 2)

3p **1c** Select all that are true:

- [ ] D is an object
- [ ] D is an array
- [ ] D is a data frame
- [ ] D is a dictionary
- [ ] D is a dinosaur

Finite Volume Methods

3p **2a** Select the correct option to fill in the "[blank]" in the following sentence:

In [blank] equations, the information travels at a finite speed.

Hint: remember that for our purposes pure advection is hyperbolic; pure diffusion can be elliptic or parabolic; and, the classification of diffusion-advection is problem-specific.

(a) Hyperbolic

(b) Elliptic or parabolic

(c) Hyperbolic, elliptic **or** parabolic

(d) Hyperbolic, elliptic **and** parabolic

(e) None of the above

3p **2b** The equation:

$$\frac{\partial \phi}{\partial t} + c\frac{\partial \phi}{\partial x} = 0$$

is solved numerically using central differences in space and the resulting ODE is solved with forward Euler in time. Which type of may occur?

(a) Numerical diffusion

(b) Numerical instability

(c) Spatial accuracy order $O(\Delta x)$

(d) All of the above

5p **2c** Recall that Gauss's theorem can be written as:

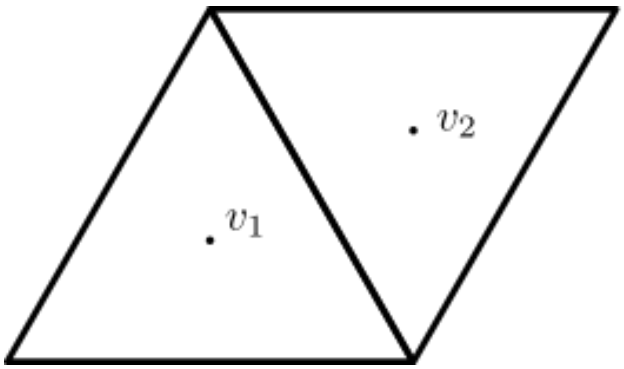
$$\int_V \nabla \cdot \vec{r}\,\mathrm{dV} = \int_S \vec{n} \cdot \vec{r}\, dS$$

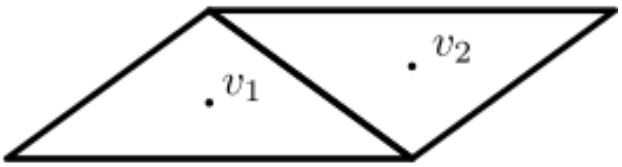
Integrate the diffusion equation:

$$\frac{\partial \phi}{\partial t} = D\frac{\partial^2 \phi}{\partial x^2}$$

over an arbitrary volume and apply Gauss's theorem to the diffusion term. There is no need to solve the surface integral.

The flux due to diffusion in orthogonal meshes, as shown below. can be approximated adequately with a central difference scheme. However, in non-orthogonal meshes, new error sources may appear.





3p **2d**

State the two main reasons why errors would appear.

Finite Element Methods

Consider the following code for the computation of one contribution to the discrete system of equations with the finite element method in 1D.

```
1  def get_element_vector(value, dx):
2      locations = [(dx - dx/(3**0.5))/2, (dx + dx/(3**0.5))/2]
3      weights = [dx/2, dx/2]
4
5      elvec = np.zeros((3,1))
6
7      for x, w in zip(locations, weights):
8          N = evaluate_N(x,dx)
9          elvec += np.transpose(N)*value*w
10
11     return elvec
```

This function is part of a finite element code that solves the PDE $a\frac{\partial^2 u}{\partial x^2} + bu = c$ for $u(x)$ with given parameters a, b and c.

3p **3a** Which of the three parameters (a, b or c) is present in this function under the variable name "**value**"?

(a) a

(b) b

(c) c

(d) The same function can be used for a, b and c.

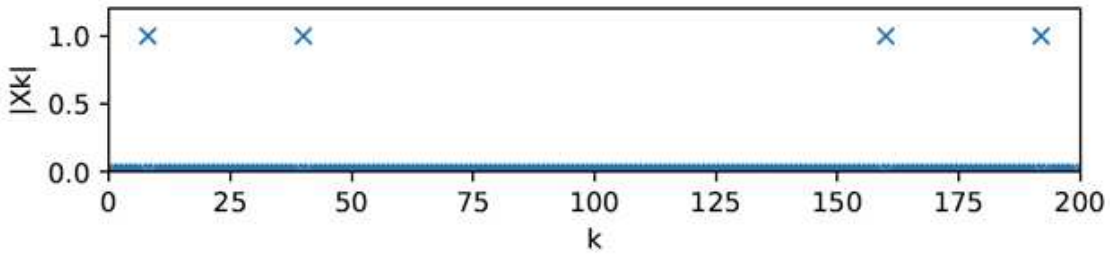
4p **3b** If at one end of the domain, no boundary condition is implemented in the finite element code, which boundary condition will be imposed by default? You do not need to write more than one equation and 1-2 sentences.

Signal Processing

We start from a signal which is the sum of two cosines, both with unit amplitude and zero phase, one with a frequency of 4 Hz, and the other with 80 Hz. The signal is sampled at $f_s = 100$ Hz, for a duration of $T = 2$ seconds.

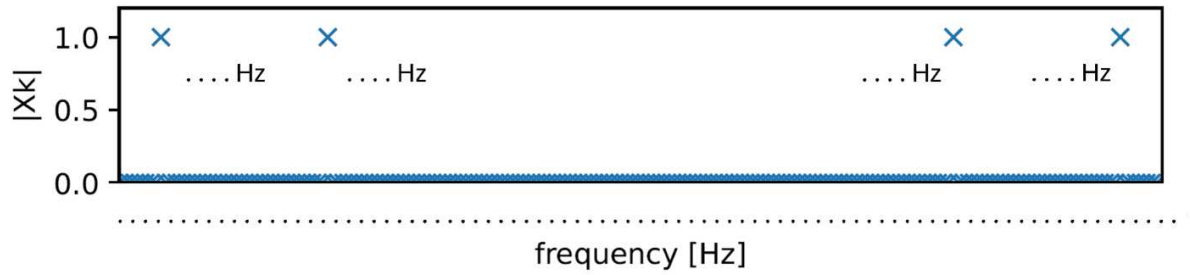
The $N = 200$ discrete time samples are input to the Discrete Fourier Transform (DFT). We directly plot the magnitude (modulus) of the output, hence $|Xk|$, multiplied by $D_t = \frac{1}{f_s}$, with $k = 0, 1, \ldots, N-1$. *(i.e. Xk=np.fft.fft(xt), and plt.plot(abs(Xk)/fs) in code).*

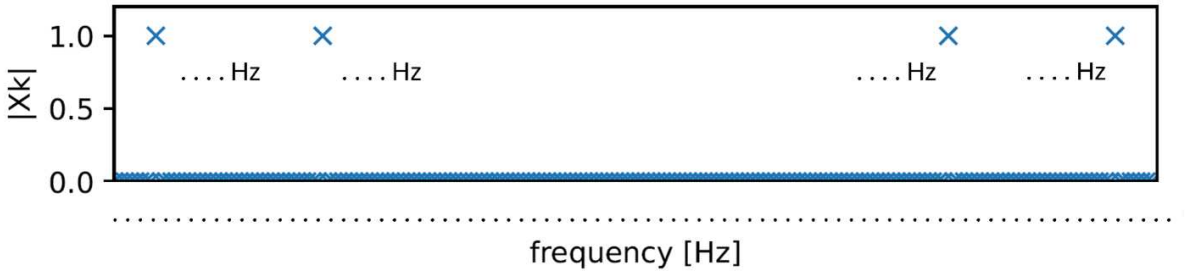
The resulting graph is shown below:



5p **4** Provide an adequate and correct labelling of the horizontal axis, such as the minimum and maximum bounds, **and** annotate the four peaks.

Please note that we have provided two images for you to use in case you need to start over. **Please cross over the image you do NOT want to submit.** If it is unclear which image you want to submit, we will only grade the bottom one.





Please add the requested numerical values on the dotted lines.

Time Series Analysis

The ice thickness at a specific location is measured on a weekly basis during 10 weeks $(t = 0, 1, \ldots, 9)$, and is assumed to increase linearly in time.

Least-squares is applied to fit a linear trend, which provides $\hat{\mathrm{x}} = [\hat{x}_0,\ \hat{v}]^T = [1.6,\ 0.12]^T$.($\hat{x}_0$ is the estimated intercept, $\hat{v}$ the estimated slope).

The estimated linear trend is then subtracted from the observations y to obtain the stationary time series:

$$S := \hat{\epsilon} = y - A\hat{x} = \begin{bmatrix} 0.23 \\ 1.27 \\ 1.16 \\ -0.52 \\ -1.90 \\ -0.50 \\ -1.60 \\ -0.62 \\ 0.04 \\ 1.23 \end{bmatrix}$$

The stochastic process underlying this stationary time series can be modeled as an AR(1) process with $\phi_1 = 0.45$.

Below you can find important formulas that you may need to use.

AR(p) process: $S_t = \sum_{i=1}^{p} \phi_i S_{t-i} + e_i$
with zero mean random noise, $\mathbb{E}(e_t) = 0$

Autocovariance function for AR(1) process: $c_r = \sigma^2 \phi^T$

Normalized autocovariance function: $\rho_r = \dfrac{c_r}{c_0}$

Prediction formula: $\hat{Y}_p = \mathrm{A}_p \hat{X} + \hat{\epsilon}_p$

Note that this information is for the next three questions on the following pages.

4p **5a** Make a sketch (no calculation needed) of the ACF for an AR(1) process with $\phi_1 = 0.45$.

4p **5b** Assume $e_{10} = 0$, what is the predicted value of $\hat{\epsilon}_{10}$?

4p **5c** What is the predicted value of the ice thickness in week 10?

Optimization

A logistics company manages the transportation of goods between two warehouses (A and B) and three distribution centers (C1, C2, and C3). The company must transport at least 500 kg of goods to C1, 400 kg to C2, and 300 kg to C3.

The transport vehicles have the following constraints:

- The total weight transported from A to all centers cannot exceed 700 kg.
- The total weight transported from B to all centers cannot exceed 800 kg.
- For operational reasons, the weight transported from A to C2 must not exceed 300 kg.

The cost of transporting 1 kg of goods between locations is given below:

From/To	C1	C2	C3
A	2	3	1
B	4	2	5

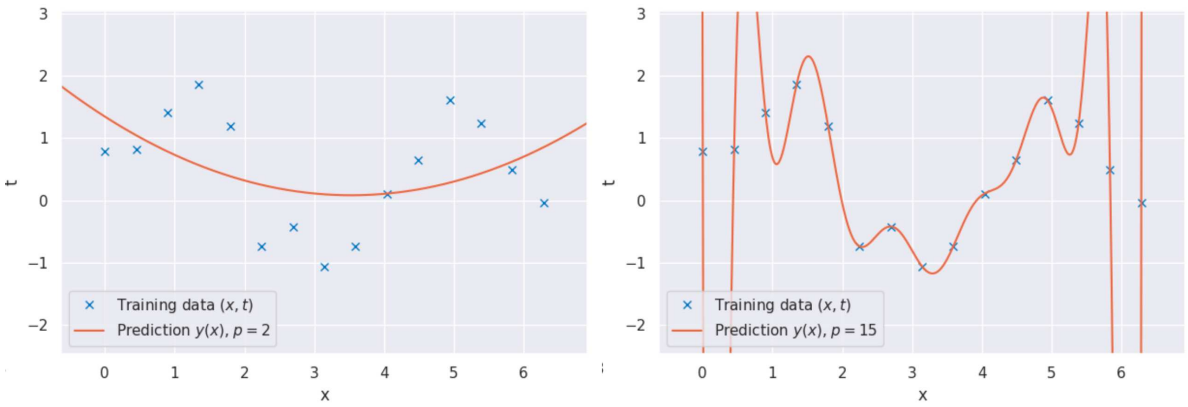
8p **6a** Formulate the problem of deciding the amount that should be transported between warehouses and centers. In your formulation please include the decision variables, the objective function and the constraints.

4p **6b** One of the distribution centers (e.g., C1) has an urgent demand for goods. Therefore you want to impose that at least 60% of the total weight transported to C1 must come from Warehouse A. Write the constraint that you need to add to the problem.

Machine Learning

3p **7a**

Consider the two following models of the form $y(x) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(x)$, where $\boldsymbol{\phi}$ is a set of polynomial basis functions up to degree p:



(a) Model A: Polynomial with degree $p = 2$

(b) Model B: Polynomial with degree $p = 15$

which are both fitted to a dataset with $N = 15$ noisy observations of target t. The prediction line shows the regression function $y(x)$. Regarding the suitability of these two models in making new predictions as part of an engineering decision-making process, which **ONE** of the following statements is true?

a) Decomposing the expected loss $\mathbb{E}[L]$ into bias, variance and noise parts, Model A would have higher variance (and therefore lower bias) than Model B

b) By adding an L_2 regularization term $\frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ to the error function of Model A, we could reach a similar level of complexity as Model B as we increase the value of λ

c) Starting with Model B and adding an L_2 regularization term, we could reach a suitable model by increasing λ until enough bias has been added to the model

d) Model B is a classic example of an overfit machine learning model. Assuming we are using a validation set to perform model selection, the overfit nature of Model B could be detected by noticing that both training and validation losses go to zero after training

3p **7b** Still regarding the two polynomial regression models above, which **ONE** of the following statements is true?

(a) Even though the regression functions have a complex shape, both A and B are still linear regression models in the original feature space $\mathbf{x}$

(b) Since the models are nonlinear, there is no closed-form (analytical) solution for $\mathbf{w}$, with SGD being therefore required for training

(c) Model B is equivalent to a neural network with a single hidden layer with 15 units (neurons) with sigmoid activation

(d) Both models assume Gaussian observation noise, which means they would be unsuitable to problems with two or more distinct ranges of t values for the same x

The information here is for us in answering 7c, which is written in the code that follows.
Now assume we would like to fit a neural network to the same dataset with $N = 15$ observations in the figure above. We opt for a Stochastic Gradient Descent optimizer and pre- and postprocess the data accordingly. Consider the following list of operations:

A) Fit input and target normalizers to the full dataset and then normalize training and validation sets separately.
B) Compute the loss over the full training and validation sets and append them to separate lists.
C) Shuffle the training dataset and split it into minibatches withN_b samples each.
D) Plot the evolution of training and validation losses against epochs.
E) Fit a single set of normalizers (input/target) to the training dataset and use it to normalize both training and validation sets.
F) Use the network to get predictions $y(x)$for all validation samples, and then plot them against the corresponding targets t.
G) Shuffle the dataset and split it into a training set and a validation set.
H) For each minibatch, compute $L = \frac{1}{N_b}\sum_{n=1}^{N_b}(t_n - y_n)^2$ and its gradient and update network weights.
I) Use the network to get predictions $y(x)$ for all validation samples, denormalize them, and then plot them against the corresponding targets t.

6p **7c** Fill in the three blanks in the code block with letters referring to the operations in the list above. Not all operations need to be used, and each blank contains more than one operation. Within each blank, make sure **the operations appear in the correct order** in the code.

Please note that we have provided two images for you to use in case you need to start over. **Please cross over the image you do NOT want to submit.** If it is unclear which image you want to submit, we will only grade the bottom one.

```
def train_model(data_X, data_t, n_epochs):
  model = init_network()

  # some operations go in this blank:
  # ------------------------

  for epoch in range(n_epochs):
    # some operations go in this blank:
    # ------------------------


  # some operations go in this blank:
  # ------------------------

  return model
```

```
def train_model(data_X, data_t, n_epochs):
  model = init_network()

  # some operations go in this blank:
  # ------------------------

  for epoch in range(n_epochs):
    # some operations go in this blank:
    # ------------------------


  # some operations go in this blank:
  # ------------------------

  return model
```

Extreme Value Analysis

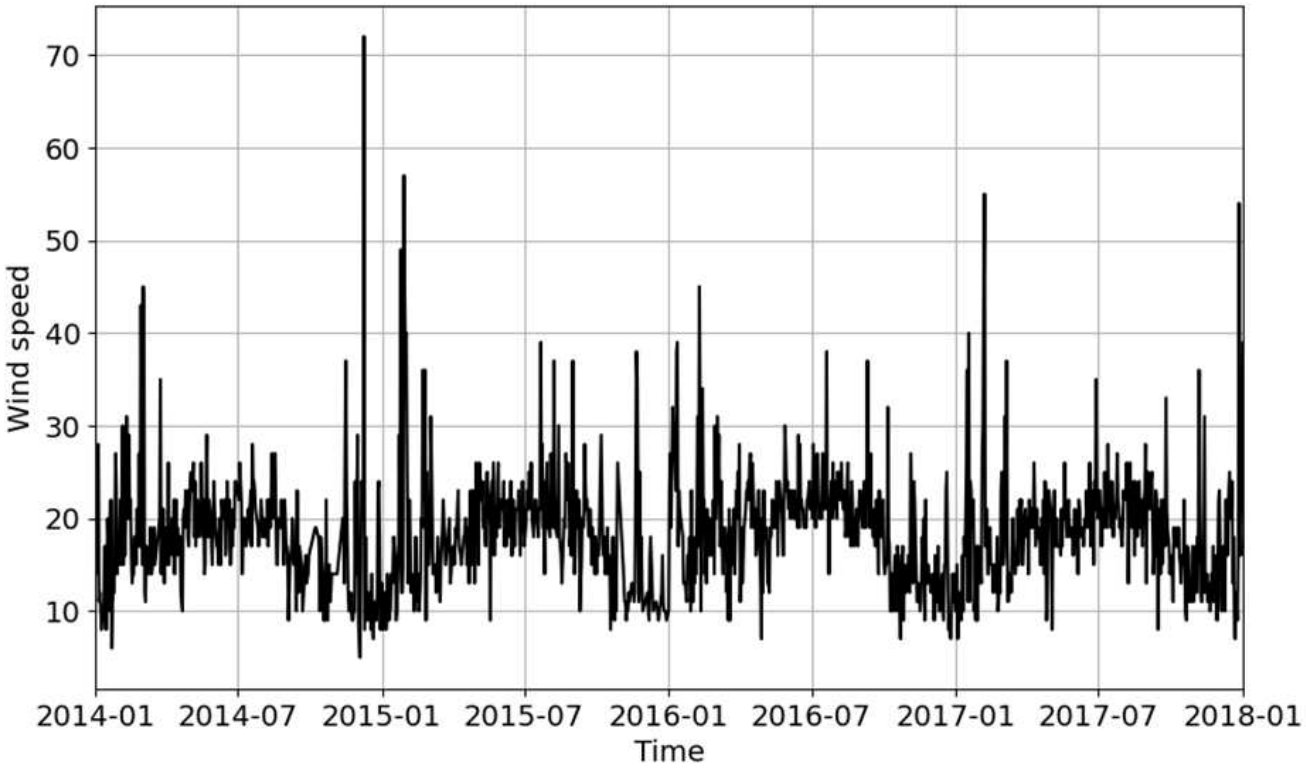
You are investigating extreme wind speeds and their impacts on buildings along the Netherlands together with a team. A colleagues studying a region close to the German border asked for your help.

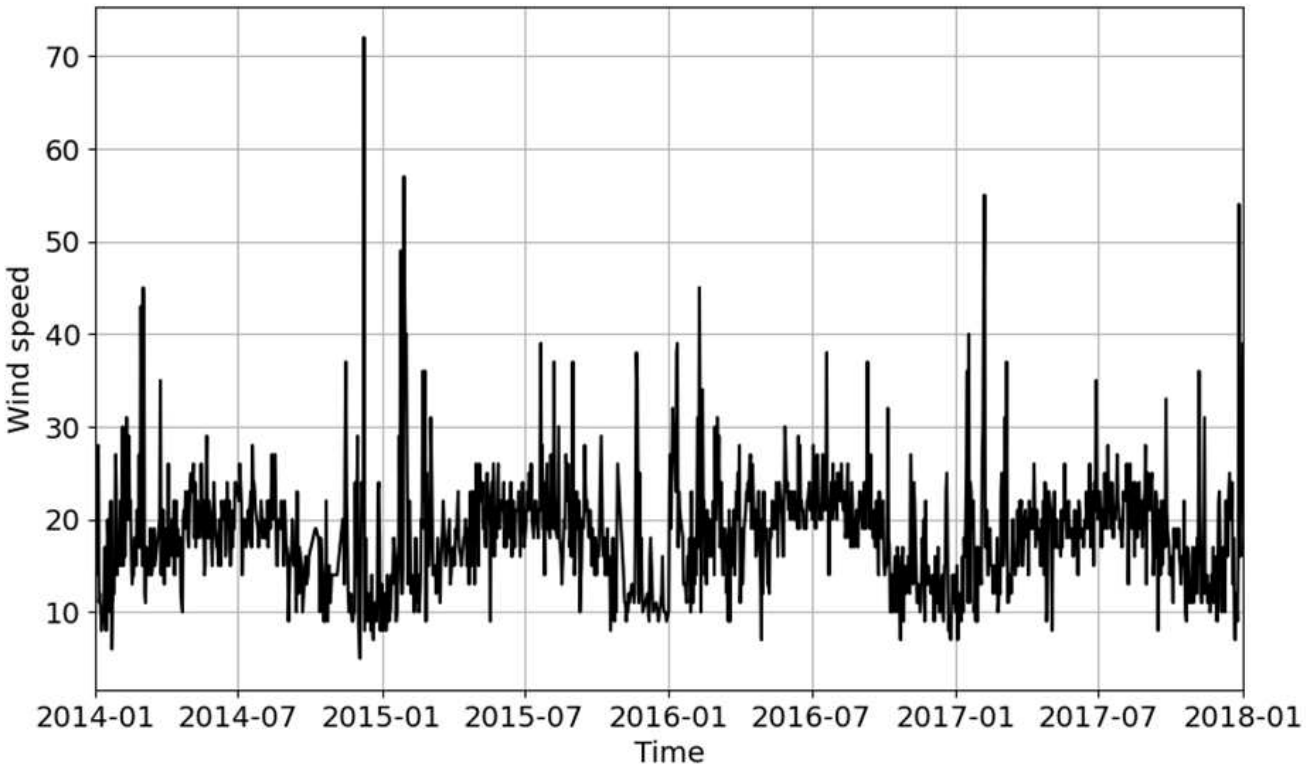
You are told that the design wind speed (W) for the buildings in that region is 40 km/h which corresponds to a return period of 50 years and that the design life of the buildings is 25 years.

3p **8a** Choose ALL THE CORRECT statements. There might be more than one.

- ☐ The expected time between two events equal or higher than 40km/h is 50 years.
- ☐ An event equal or higher than 40km/h will not be observed in the next 50 years.
- ☐ An event equal or higher than 40km/h will not be observed in the next 25 years.
- ☐ Every year $P\left[W > 40\right] = \frac{1}{50}$
- ☐ Along a design life of 25 years, $P\left[W > 40\right] = \frac{1}{50}$
- ☐ The probability of $P\left[W > 40\right]$ along the design life of 25 years is higher than the yearly probability of $P\left[W > 40\right]$

4p **8b** Given the following figure of the time series between 2014 to 2017, how would you find the yearly maxima?
First, describe ONLY the sampling process briefly in the space below; **second,** use the plot provided to sketch it. Please note that we have provided two images in case you need to start over. **Please cross over the image you do NOT want to submit.** If it is unclear which image you want to submit, we will only grade the bottom one.


Wind speed
70
60
50
40
30
20
10
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Time


Wind speed
70
60
50
40
30
20
10
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Time

6p **8c** Another colleague asked you for your help as probability expert. They have already performed yearly maxima on the observations of wind speed (km/h) and have fitted a Generalized Extreme Value distribution (GEV) so:

$$W \sim \text{GEV}(\mu = 25, \sigma = 10, \xi = -1)$$

The GEV probability density function is given by:

$$g(x) = \frac{1}{\sigma}\left[1 + \xi\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}(\xi+1)} \exp\left(-\left[1 + \xi\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right)$$

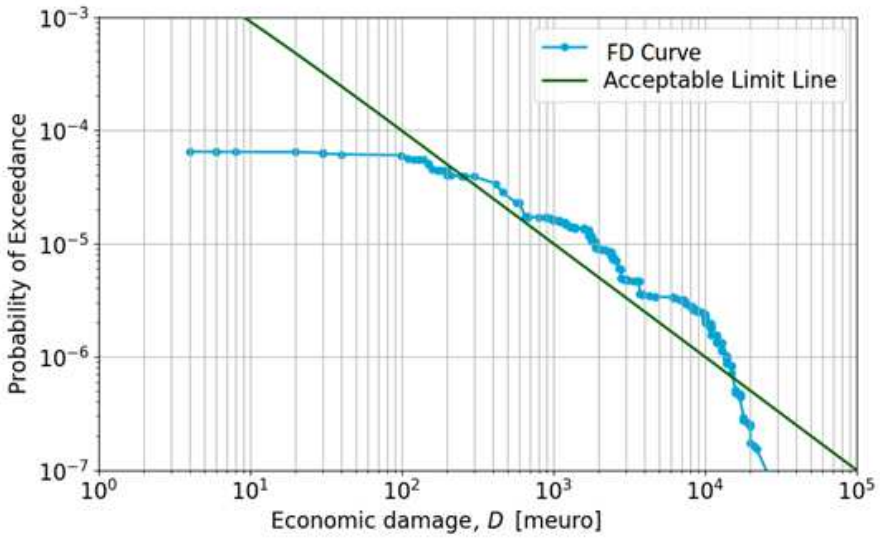
and the GEV cumulative distribution function is given by:

$$G(x) = \exp\left(-\left[1 + \xi\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right).$$

Compute the wind speed associated with a return period of 50 years.

Risk & Reliability

When answering the following 2 questions , consider the figure here, which illustrates a risk analysis for a flood protection area.



3p **9a** Which of the following would be appropriate ways to *meet* the safety requirement? (more than one can be selected)

- ☐ Remove people from the area.
- ☐ Make the flood protection stronger
- ☐ Convince policy makers to allow for a higher chance of fatalities
- ☐ None of the above

3p **9b** Identify the false statement in the list below (there is only one). The statement should complete the sentence: *The FD curve for the flood protection system...*

- (a) ...quantifies probability of system performance
- (b) ...quantifies probability of component failure in the system
- (c) ...illustrates consequences of system failure
- (d) ...can be used to compute expected value of consequences
- (e) ...illustrates exceedance probability

Use the grid below if you run out of space for any exercise.
In that case, please indicate so at the original answer field.

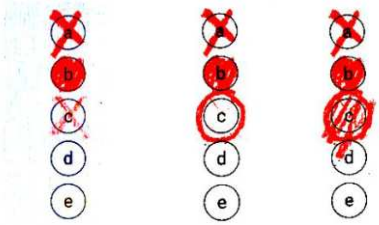
Exam Overview and Multiple Choice Examples

The table below gives an overview of the questions to help you plan your time during the exam.

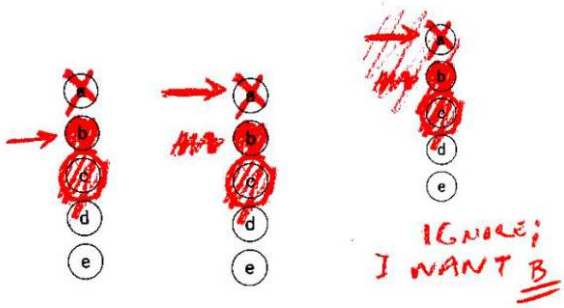
No.	Topic	Number of Sub-parts	Points
1	Programming	3	9
2	Finite Volume Methods	4	14
3	Finite Element Analysis	2	7
4	Signal Processing	1	5
5	Time-series Analysis	3	12
6	Optimization	2	12
7	Machine Learning	3	12
8	Extreme Value Analysis	3	13
9	Risk & Reliability	2	6
Total			**90**

In case you want to correct your answer for a multiple choice question put an ARROW in front of your final answer. If you also make a mistake with your arrow, write a clear message on the page. Here are a few examples:

Examples of UNCLEAR multiple choice response:



Examples of CLEAR multiple choice response:



Answer: B Answer: A Answer: B

Multiple choice examples.