Water Resources Engineering
Department of Civil and Environmental Engineering
Helsinki University of Technology

Integrated Catchment Assessment and Management Centre
Fenner School of Environment and Society
The Australian National University

COMPUTATIONAL METHODS FOR WATER RESOURCE ASSESSMENTS: AN EXERCISE KIT

Edited by

Teemu Kokkonen
Anthony J. Jakeman
Harri Koivusalo
John P. Norton

ISBN 978-88-903574-0-4

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

During the compilation of this exercise book the first editor was funded predominantly by the Academy of Finland, whose support is greatly acknowledged.

We are grateful to Professor Pertti Vakkilainen's for his support and encouragement throughout the writing process of this book. Professor Tuomo Karvonen's example in constructing exercises that clarify how simulation models for water resources assessments are built has been an important source of inspiration for this book.

Contribution of Jason Pulkkinen and Leena Stenberg in preparing illustrations of this book is greatly acknowledged.

We are grateful to William Francis for designing the cover of this book.

1 INTRODUCTION

T. Kokkonen

Motivation for this book lies in the firm belief of the value in learning by doing. The book attempts to impart an understanding of how computational methods can be used in water resource assessments by providing exercises related to different aspects of the cycle of water in the environment. Many of the exercises are built on real data measured in the field. This approach hopefully increases the readers' interest in the subject as well as their appreciation of uncertainties, and even sheer errors, appearing in measurements.

The intent of the present book is not to provide a detailed coverage of the various methods and topics. Instead, it is presumed that before embarking on the exercises readers have learnt the relevant theory by attending lectures or by studying more comprehensive text books. But after readers have attained a sound knowledge of the basic theory, the present book is intended to be self-sufficient in giving a brief, concentrated dose of theory that is necessary for tackling the exercises.

The exercises are developed to encourage readers to explore what happens 'behind the scenes' in environmental simulation models. Hence in many exercises the computational procedures are built almost from scratch before applying the model to a particular problem. It is believed that gaining understanding of how computational methods used in environmental assessments actually work not only promotes readers' potential to develop their own applications, but also increases their skill to assess critically existing environmental simulation models.

This book is distributed free of charge. Also, the only proprietary software required to solve the exercises is Microsoft Excel, which is typically available to most students. The Geoinformatica GIS software that is used in those exercises involving geospatial computations is free and open source software and its installation package is bundled with the distribution package for this book. **Note Jan 30, 2020**: Geoinformatica is no longer included in the distribution package of this book (see Section 12).

A majority of the exercises come as Excel worksheet templates where the solutions can be in-filled according to the instructions given in this book. In addition, Excel worksheets containing technical model solutions are also included in the book distribution package. The term 'technical' means here that, while a model solution shows one way of building the computational procedure in question, it does not include answers to questions that address applying the procedure. Technical solutions are provided to help readers deal with situations where progression with the exercise is threatened such as when a certain step in solving the exercise remains unclear. Then the reader can look up a solution to the problem blocking their way forward and then carry on with solving the exercise. This should diminish frustrations that easily surface when not understanding a single detail prevents from completion of the exercise. The access that is provided to model solutions is also intended to facilitate use of this book for independent learning as readers can compare their solutions to the suggested ones. Hence some confidence in the learning results can be attained without having a teacher available to comment on the solutions. The increased confidence that students gain by having the model solutions available calls for an increased responsibility of the students in their own learning. Fortunately it seems that students are willing to take responsibility if teachers dare to shift it onto them. But of course when this book is used in a course belonging to an obligatory curriculum, teachers should also assess to which extent students have understood the lessons of the exercises.

It is noteworthy that symbols used in this book are unambiguous within chapters. It is quite possible, however, that the same symbol in one chapter denotes a different variable in another chapter. The units of different variables appearing in the book are, for the most part, indicated explicitly in brackets. Readers are encouraged to make a good habit of always checking that the units in their calculations match, i.e. that the unit on both sides of an equation is the same. For example, computing the velocity [ms^{-1}] by dividing distance [m] by time [s] yields the unit of velocity [ms^{-1}] on both sides of the equality sign.

We hope that you enjoy the exercises in this book and that your learning profits greatly from them. Please feel free to contact me (firstname.lastname@tkk.fi) if you have suggestions for improving the exercises.

2 SOFTWARE INSTRUCTIONS

T. Kokkonen, A. Jolma

2.1 Excel

The instructions on using Microsoft Excel given below are based on Microsoft Office versions 2000 and 2005. Many of the listed functionalities are accessed in a different way in Microsoft Office 2007. The users of Microsoft Office 2007 are encouraged to use the Help system of the software in parallel with the instructions given here.

NAMES (NAMED RANGES) IN EXCEL

1. Why attach names to cell ranges?
   - This often makes worksheet functions referring to these ranges more readable.
   - If you do not wish to attach names, you can also refer to cell ranges by selecting the range with the mouse, or by typing in the worksheet coordinates of the range to be referred to.
2. How to name a cell?
   - Place the cursor on the cell (or range of cells) that you wish to give a name.
   - Click the mouse in the *Name Box* in the upper left corner (the white box that shows the coordinates of a cell, e.g. A1.
   - Type in the name and hit Enter.
3. How to use the name in a worksheet function?
   - Say you gave a cell the name 'tempK', now if that cell contains a temperature value in Kelvin, you can type in another cell '=tempK – 273' to convert that temperature to ºC.
4. How to delete a name?
   - If you need to delete a name, go *Insert->Name->Define*, select the name to be deleted, and hit Delete.
5. How to ascertain which names are in use, and what they refer to?
   - In the exercise workbooks there is often in the beginning a list of names that have been predefined in that sheet.
   - If you go *Insert->Name->Define*, you obtain a list of all names (in all open workbooks) currently in use.
     - To see what a name refers to, select the name from the list and click in the *Refers to* box.
   - When you select a range that has a name, the name appears in the *Name Box* in the upper left corner.
6. Eligible names
   - Some names (like A1) cannot be used for named ranges because they already have a meaning in Excel, e.g. A1 refers to the cell with the coordinates A1**.**

MATRIX FUNCTIONS IN EXCEL

1. Select an area that occupies the size of the resulting matrix.
2. Type in the matrix function and an opening parenthesis – '(' .
3. Give the argument(s) for the function by selecting them with your mouse or by typing them in. When there is more than one argument, each needs to be separated by either a comma or a semicolon (depending on your Windows configuration for a list separator).
4. Type in the closing parenthesis – ')' – and accept the formula with CTRL-SHIFT-ENTER

SOLVER

1. How to open Solver in Excel?
   - Select *Tools* -> *Solver*.
   - If you cannot find it you need to select *Tools* -> *Add-Ins* and to tick *Solver Add-In* from the list.
2. How to set the function to be optimised?
   - Provide in the *Target cell* box the cell containing the function to be optimised by selecting it with mouse or by typing in its worksheet coordinates.
   - The *Equal to* radio buttons allow you to choose whether you want to maximise the objective function, minimise the objective function, or try to set the objective function to a given value.
3. How to set the variables to be adjusted in optimising the function?
   - Provide in the *By changing cells* box the cells that are to be adjusted by selecting them with the mouse or by typing in their worksheet coordinates.
4. How to apply constraints?
   - Click the *Add* button on the side of the *Subject to the Constraints* box, and give the constraint using the appearing dialog box, e.g. A1 >= 10

ITERATION

1. How to enable iteration?
   - Select *Tools* -> *Options* -> *Calculation*
   - Check the *Iteration* check box.
   - You can modify the criteria when iterations are terminated. In the *Maximum iterations* box the maximum number of iteration rounds can be set, and in the *Maximum change* box the maximum change between two iteration rounds can be inserted

INDIRECT WORKSHEET FUNCTION

You can use a value from another cell in the cell reference with aid of the INDIRECT worksheet function. For example, see the figure below illustrating how the date in column B and row 11 is referred to with aid of the INDIRECT function.

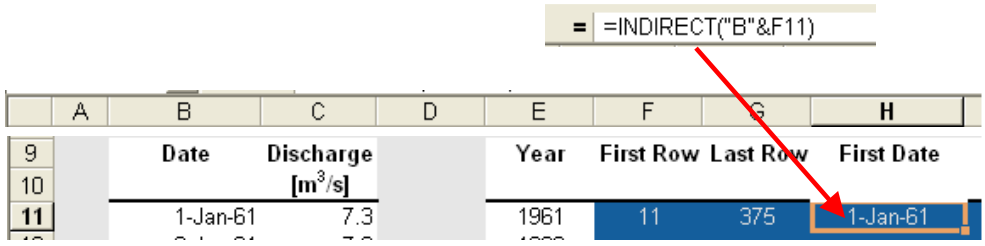

Figure 2.1. An example of using the INDIRECT worksheet function.

2.2 Geoinformatica

Geoinformatica is a GIS (Geographic Information Systems) software that builds on several free and open source software packages. Its installation package is bundled with the distribution package of this book. The documentation given below does not attempt to provide a full manual for the software. But it hopefully succeeds in providing sufficient information to get readers through those exercises where Geoinformatica is needed. As explained in the following section, Geoinformatica utilises the Perl programming language as the engine for accessing its geospatial tools. Documentation on Perl commands available in Geoinformatica is attached to the Geoinformatica distribution package, and it is available through the start menu (see Section *Install and Run* below).

Geoinformatica and Perl

Geoinformatica is dependent on the Perl programming language and includes a Perl command interpreter for accessing a set of geospatial functions. You need to know very little about programming in Perl to be able to work through the exercises of this book, but please make sure you know the following.

Variables need a prefix in front of their name

Whenever you use an existing variable, or introduce a new one, a prefix is required in front of a variable name. For scalar variables the prefix is the dollar sign ($), and for array (or list) variables it is the 'at' sign (@). In the exercises of this book you deal predominantly with scalar variables, so it (almost) suffices to remember that you always need to write the dollar sign in front of a variable name.[1] For example, if you open a grid called *test* and want to refer to it, you need to write $test. Interested readers can refer to http://perldoc.perl.org/perlintro.html#Perl-variable-types for a brief introduction to variable types used in Perl.

Methods of grid objects are accessed using the arrow (->) sign

What follows is an extremely short (but sufficient for the purposes of this exercise book) introduction to object oriented programming. An object contains both data and algorithms. Algorithms are typically referred to as methods, and in Perl they are accessed using the -> sign. For example, in Geoinformatica all grids are objects that contain the following data: cell values, number of rows and columns, cell size, data type, and bounding box of the grid in map coordinates. Grid objects also have methods for numerous operations, for example for computing the mean value over all cell values. If the name of a grid is *test* and you want to store the resulting mean in a variable called *testmean*, you would write

```
$testmean = $test->mean()
```

The empty parentheses mean that this method does not require additional arguments. Take a look at the manual entry for the mean method, which you find in the Perl modules documentation under *Classes -> Geo::Raster::Global -> mean*. You will see that the mean method calculates the mean of all grid values, and returns the mean as a scalar variable.

If you wanted to set in the *test* grid the cell value at the 10th row and 14th column to 7, you would write

```
$test->set(10, 14, 7);
```

Now the *set* method needs as arguments the row and column indices, and the new value. Please see the manual entry for the *set* method (*Classes -> Geo::Raster -> set*). The set method takes three (scalar) arguments, as listed above, and returns nothing (void).

Hash data type

A hash is a table where the table values are associated with key values. The key can be any integer or a string. For example, a hash could contain populations of cities where the city name is used as a key (Table 2.1).

Table 2.1. Key- value pairs for a hash containing city populations.

Key	Value
London	7500000
Paris	2200000
New York	8100000
Tokyo	12600000

[1] The only exception to using the $ sign as the prefix is those Geoinformatica methods that return a *list*. Then you should use the @ sign as the prefix. Often you find that the returned variable is a grid (Geo::Raster in the manual entry). This is a reference to a grid object and it also takes the $ sign as the prefix.

Some of the Geoinformatica methods return hashes. Or to be more precise, the return type is typically a reference to a hash. A reference is a scalar variable and hence the prefix for a hash reference is again the dollar sign. There is a very simple way to print out the contents (i.e. the key-value pairs) of a hash. If the hash reference variable is called *test*, just write

```
p($test)
```

to print out the hash contents to the output window.

INSTALL AND RUN

You can install Geoinformatica for Windows by simply running the Geoinformatica-2009-05-18.exe located in the *Geoinformatica* directory of the book distribution package. The installation package will guide you through the installation.

Run Geoinformatica from the start menu, or by double-clicking the Geoinformatica desktop icon. Documentation of Geoinformatica is also available through the start menu.

BASIC FUNCTIONALITIES

Opening data

You can open grid (raster) data to Geoinformatica by clicking *Open raster*, browsing to the directory where your grid data reside, selecting the grid data file (all grid data included in this book have the *.tif* extension), and clicking *OK*. **Important note!** Before starting to use the grid data in your analysis you should right-click on the name of the data set visible on the left, select *Clip…* from the menu that appears, select <*self*> in the *Clip to raster* drop-down box, and finally click *OK*. This guarantees that the entire data set is loaded to the RAM (Rapid Access Memory) of the computer and is hence available for the analysis. Geoinformatica is able to open and visualise very large grid data sets by accessing them from the hard disk and plotting them on the screen with a lowered resolution (i.e without having all the data stored in the RAM). Doing the clip for a data set that is too large for the amount of RAM available on the computer will almost surely result in an 'out of memory' error message. But this should not be a problem for the relatively small grid data sets used in the exercises of this book.

Vector data are opened similarly to grid data by clicking *Open vector*, browsing to the directory where the vector data are located, selecting the desired vector layer in the *Layer* column, and then clicking *OK*.

Saving data

There are several ways to store grid data sets on the disk. You can right-click on the grid layer name, select *Save…* from the menu that appears, and then give the filename and select the folder. Clicking *OK* stores the data with the given filename in the selected folder using the bil/hdr format. Another way to store grid data is to click *Save rasters* in the main menu bar, which opens a dialogue box for selecting a folder. After having selected a folder click *OK*, and all raster layers loaded to RAM and visible on the left will be stored in the selected folder. A third way to save a grid is to invoke its save method from the command line. If the name of the grid is *grid1* and you wish to save it in the directory *C:\ Data* you can write in the Perl command interpreter of Geoinformatica (white bar at the bottom)

```
$grid1->save('C:\ Data\grid1')
```

Geoinformatica can also save vector data. This can be achieved by first selecting a set of features (right-click on the map, activate the *Select* mode, and drag with the mouse) and clipping the selected features then into a new vector file (right-click on the name of the vector layer and select *Clip…*). Saving vector data, however, is not necessary in the exercises of this book.

Zooming and panning

The handiest way to zoom in is to click and drag the mouse on the map to select the region where to zoom to. Another way is to right-click on the map and select from the appearing menu *Zoom in*.

To zoom out you can either select *Zoom out* from the menu available by right-clicking, or you can click *Zoom to all* in the main menu bar, which zooms to the full extent all data layers loaded into the map.

Other useful ways of zooming include zooming to the extent of a single layer, or zooming to the previous extent. Right-clicking on the name of a layer allows you to select from the appearing menu *Zoom to*, which zooms to the extent of that layer. And right-clicking on the map opens the menu where you can select *Zoom to previous*, which zooms to the extent that was visible just before. Zooming to the previous extent can be repeated, which allows you to reverse back to the extent that was active several zoom actions ago.

To activate pan mode, right-click on the map and select *Pan*. Now you can pan the map by dragging it with the mouse. If you wish to start zooming again you need to right-click on the map and select *Zoom* mode. The zoom mode is the default mode that is active when you start Geoinformatica. The currently active mode is indicated by *x* in the menu that opens by right-clicking on the map.

Setting no-data value

The term no-data value refers to a special grid cell value that is assigned to those cells whose values are unknown or unimportant. It needs to be selected in such a way that it cannot be confused with a valid data value. For example, should you have elevation data from low-land areas, zero would be a bad choice for the no-data value, because a part of your terrain may actually reside at sea level, i.e. at zero elevation. But -999 (the data in metres) would be a good no-data value as it cannot be mixed with any valid data values.

Assigning a no-data value serves two purposes. First, grid cells having a no-data value are not included in any calculations. So unimportant regions can be ignored by assigning them with a no-data value. And second, in Geoinformatica grid cells having a no-data value are always transparent, which can increase clarity in visualizing several grid layers on top of each other.

You can change the no-data value by right-clicking on the name of a grid data layer, selecting *Properties…*, and typing the new value in the *Nodata:* box.

Colour schemes and symbols

The colour scheme of a grid data set is controlled by right-clicking on the name of the data set and selecting the *Colors…* entry from the menu. From the *Palette* drop-down list you can set the colouring palette for the current data set. For continuous change of shades you can apply, for example, a *Rainbow* (in colour) or *Greyscale* (shades of grey) palette. Geoinformatica can determine the range of values in the current data set when you click the *Get range* button, but leaving the values at zero instructs the software to always adjust to the minimum and maximum of the visible part of the grid. After the range has been set, click *Make legend* for generating the legend for the selected data set, range, and colouring palette. Clicking *OK* applies the selected colouring scheme to the data. The *Single color* palette applies one colour to all cells containing valid data (i.e. anything else but no-data) cells. This is a good way of visualizing binary data, which could for example be a catchment area where a value (e.g. 1) indicates that the cell belongs to the catchment area, and another value (e.g. no-data) means that the cell resides outside of the catchment area. To set the one colour for the *Single color* palette, select the line showing the color (by default white) along with its red, green, blue, and alpha values and then click the *Color* button. Now click the pipette icon and then use the pipette to select the colour you require from the circle. You can also adjust the lightness of the colour from the inner triangle. Clicking *OK* in the colour selector and the colour scheme window applies the *Single color* palette to the present data set. *Color table* and *Color bins* palettes are generated analogously to the *Single color* palette, but now you can define several colors for a set of values (*Color table*) or a set of value bins (*Color bins*). *Red channel*, *Green channel* and *Blue channel* are meant for visualizing e.g. multiple channel satellite images and they are not well suited for the data given in the exercises of this book.

Colour scheme for a vector data set is generated in a similar manner as for grid data. For other than *Single color* palette, however, you can select the field on whose value the colour scheme is based from the *Color is based on:* drop-down list. It is also noteworthy that the border colour of polygon data is controlled from the *Properties…* entry of the menu that opens by right-clicking on the name of a vector layer. The border colour is black by default, which can cause polygon data to appear all black, independent on the selected colour scheme, unless you zoom close enough. To avoid this you can uncheck the *draw border on polygons* from *Properties…*, or alternatively change the border colour from the same place.

Transparency of both grid and vector data is controlled by setting the alpha value from the *Properties…* dialogue that is accessed by right-clicking the name of the data set. An alpha value of 255 indicates fully opaque colours, and setting the alpha value to zero makes the layer totally transparent (i.e. invisible). A value between these two extremes makes the layer partially transparent, which can be useful as it allows you to view multiple grids at the same time.

In addition to choosing different colours for grid and vector data you can also use symbols for visualization. The control of symbols is accessed by right-clicking on the name of a data layer and selecting the *Symbol…* entry. This functionality becomes useful in Exercise 12.1 for studying the flow network as interpreted from a digital elevation model. Please note that the symbols become visible only when you zoom in close enough to the data set having symbols.

Controlling the map view

The data layers are shown in the map view in the same order as their names appear in the panel on the left side of the map view. You can change the order by right-clicking on the name of a data layer, and selecting then either *Up* or *Down* from the menu. You can hide a data layer by selecting *Hide* from the same menu, and then make it visible again by selecting *Show*. A cross (x) on the right of the name indicates whether the data layer is currently shown. If you wish to unload a data layer from the map, select *Remove*. Note that the remove command does not warn you when you delete unsaved grids.

Using Perl command interpreter

The Perl command interpreter is the white bar at the bottom of the Geoinformatica window. Just click in there and you are ready to type commands in Perl. Recall the earlier example of computing the mean value over all grid cells. If you have a grid data set called *test* loaded into the map, you can compute its mean and store the result in a variable called *testmean* by typing in the Perl command interpreter

```
$testmean = $test->mean()
```

The output window is opened by selecting *Output* from the main menu bar. All output generated by commands given in Perl appear in the output window. In this case you can see the result by issuing for example the command

```
print "$testmean\n"
```

or

```
p($testmean)
```

and opening the output window.

Status bar

The status bar below the Perl command interpreter shows the following information for grid data: the map coordinates (*x* and *y*, origin is at lower left corner), the row and column coordinates (*i* and *j*, origin is at top left corner), and the grid value of that location where the cursor (tip of the arrow) resides. For vector data only map coordinates are shown. Moreover, the currently active operation mode (e.g. zoom or pan) is also indicated.

BUGS AND SUPPORT

A good place to start looking for more information about Geoinformatica is http://trac.osgeo.org/geoinformatica/. Geoinformatica is not an official OSGeo (Open Source Geospatial Foundation) project, but it relies heavily on GDAL, which is an official OSGeo library project. On that page there are links to the Geoinformatica bug and issue tracker, user-oriented wiki pages, the source code, binary releases, documentation, and information on how to contact the developers and other users.

3 MODELLING CONCEPTS

T. Kokkonen, A. J. Jakeman, J. P. Norton

3.1 Calibration

Model calibration means adjustment of model parameters in such a way that the model output matches the measured data as closely as possible, in accordance with some measure or metric (such as sum of least squares error, discussed later in Section 4.2). For an example, see Figure 3.1 where black dots represent the measured relationship between *x* and *y* variables. Now if it is assumed that the relationship between *x* and *y* is linear and a line is to be selected as the model for describing how *y* depends on *x*, model calibration means selecting those slope and intercept values that the line fits to the measured points in some best possible way. From the three depicted lines with different parameter (i.e. slope and intercept) values, line C clearly resembles most closely the measured relationship. Therefore, given these alternatives any sensible calibration method would pick slope and intercept values, that yield something close to line *C,* as the calibrated parameter values[2]. The actual values would depend on the specific nature of the metric.

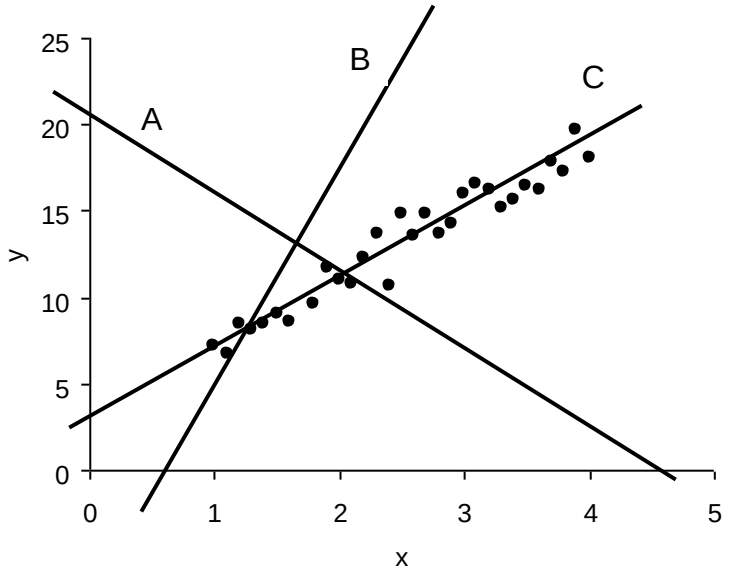

Figure 3.1. Measured relationship between *x* and *y* variables, and three alternative lines (A, B, C) for describing the assumed linear relationship between the two variables.

[2] If the observations (*y* here) are taken as resulting from observation errors adding to the values given by the assumed relationship, the situation is less straightforward. The parameter values which minimise the misfit between the y values and the outputs of the relationship (the line in this example) may then not be the same as the values minimising the error in those parameters. This is so when there is some systematic relation among the observation errors or between them and the explanatory values (of *x* here).

3.2 Validation

The term model validation typically refers to testing performance of the model using independent data, i.e. data that have not been included in the model calibration. This is a very important step in model development, and unfortunately it is not always given the attention that it deserves. Figure 3.2 depicts again the measured relationship between x and y indicated with black dots. The data where *x* is smaller than or equal to four were used in calibration and line *C* was selected as the model that best matches the measurements (see Figure 3.1). Now when model performance is compared against independent data where *x* is greater than four it can be concluded that the model fit is significantly inferior in validation phase than in calibration phase. Parameter values of the line would be different if validation data had been included in the calibration, but then no data would have remained for testing the model performance. Moreover, the validity of assuming a line to be capable of representing the relationship between *x* and *y* in the validation phase can be questioned. The curvature visible in the *x*-*y* relationship indicates that *y* depends nonlinearly on *x*, and then the line would be an incorrect model structure. In conclusion, the example shown in Figure 3.2 demonstrates that the information content in the data may not always be sufficient to identify model parameter values and model structure correctly. If model performance drops significantly in validation it is advisable to study why this might have happened and possibly revise or reject the model. A successful validation test, on the other hand, increases confidence in the selected model structure and parameter values although it cannot confirm for certain that they are the best, or correct.

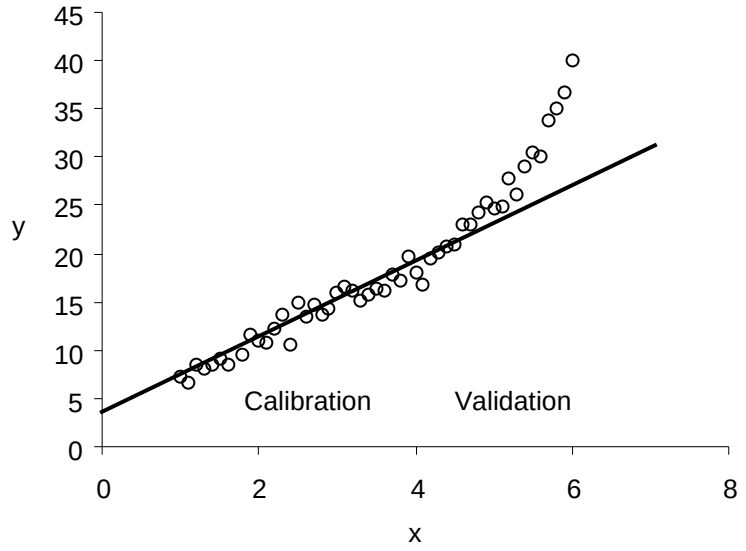

Figure 3.2. Measured relationship between *x* and *y* variables, and a line whose parameters have been calibrated for those data where *x* is smaller than or equal to 4. Model performance is checked for data where *x* is greater than 4.

We now proceed to present some basic measures that are often used to evaluate a model or test its performance

4 STATISTICAL METHODS

T. Kokkonen, A. J. Jakeman, J. P. Norton

4.1 Statistical concepts

MEAN, EXPECTED VALUE AND MEDIAN

The mean is a measure that describes where the centre of a set of values, from x_1 to x_N, is located. It can be estimated from

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{4.1}$$

where N is the number of values in the set. Equation (4.1) yields a sample estimate of the mean based on a sample size of N. The quality of the estimate increases with an increasing sample size, but it generally differs from the actual (ensemble) mean whose value is often unobservable. The expected value is in statistics a synonym to mean. For a discrete random variable x it is defined as

$$E(x) = \sum p_n x_n \tag{4.2}$$

where E is the expected value operator, p_n is the probability of the n^{th} outcome of x, and x_n is the value of the n^{th} outcome of x. The expected value of rolling an ordinary six-faced die is hence

$$E(x) = \frac{1+2+3+4+6}{6} = 3.5 \tag{4.3}$$

Equation (4.1) could be used to deliver an estimate of the expected value of rolling the die by repeating the rolling N times and inserting the outcomes in (4.1). But it is noteworthy that use of the sample mean in place of the actual mean generally affects the analysis. However, often this is all we can do.

Another measure of the centre, the median, is the middle-ranking value in the data set. Unlike the mean, it is insensitive to the size of any extreme values in the set. This can sometimes be useful, especially when a small proportion of the values are 'bad data', e.g. subject to large measurement errors. Let us consider two data sets: Set1 = {1, 2, 3}, and Set2 = {1, 2, 102}. The median is 2 for both sets, whereas the mean is 2 for Set1, and 35 for Set2.

VARIANCE AND STANDARD DEVIATION

The variance σ^2 is a measure of spread of values around their mean. It can be estimated from[3]

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_i - \bar{x}\right)^2 \tag{4.4}$$

Figure 4.1 depicts two sets of 31 values where both have the same mean of 2, but the variance is greater in Figure 4.1b. Standard deviation σ is defined to be the square root of the variance.

[3] It is noteworthy that equation (4.4) yields a sample estimate of the variance, just like (4.1) gives a sample estimate of the mean.

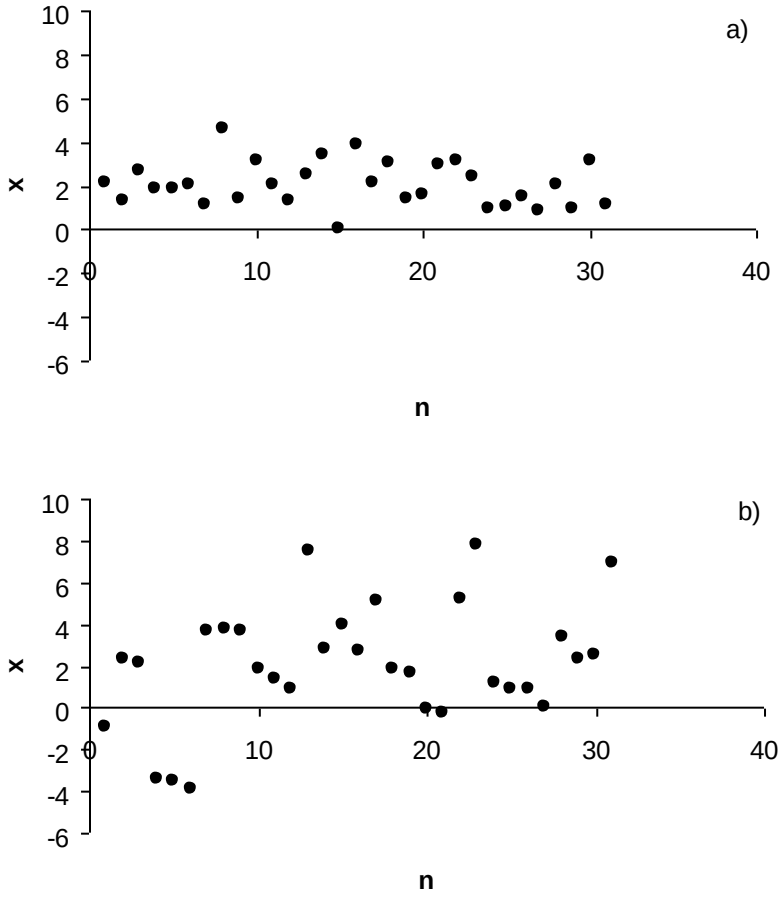

Figure 4.1. Two sets of 31 values (x_n is the value of the n^{th} observation). Both sets have the same mean, 2, but the variance is greater in b) than in a).

COVARIANCE AND CORRELATION

The covariance $\sigma_{x,y}$ is a measure of the extent to which two variables (x and y) move together. It can be estimated from[4]

$$\sigma_{x,y} = \frac{1}{N}\sum_{n=1}^{N}\left[\left(x_n - \bar{x}\right)\left(y_n - \bar{y}\right)\right] \tag{4.5}$$

A positive covariance value indicates that an increase in x is likely to be accompanied by an increase in y, whereas a negative covariance suggests that an increase in x is likely to be accompanied by a decrease in y. The magnitude of covariance is dependent on the magnitude of the values of x and y. This makes it harder to interpret if a given covariance value is large or small. For this reason, it is convenient to standardize the covariance to range between –1 and 1. Then it is called the correlation $\rho_{x,y}$, defined as

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \tag{4.6}$$

where σ_x and σ_y are standard deviations of x and y, respectively. Figure 4.2 depicts two datasets where the x and y variables have a positive (a) or a negative (b) correlation.

[4] This is also a sample estimate. See footnote 3 on page 11.

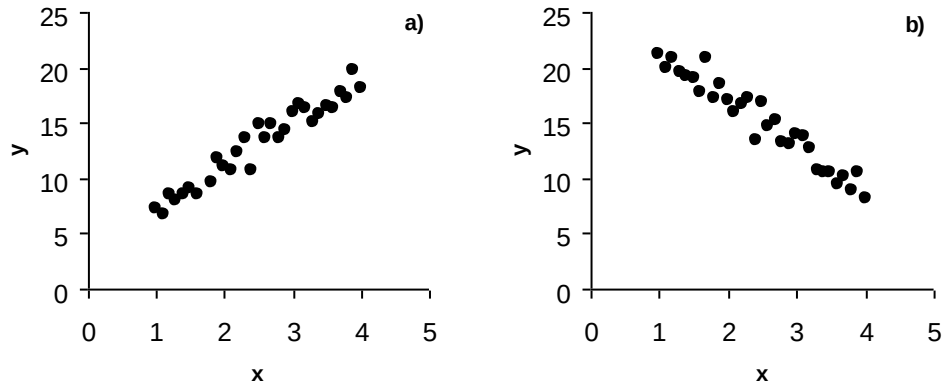

Figure 4.2. Positive (a) and negative (b) correlation between *x* and *y*.

BIAS

Assume that the points shown in Figure 4.2a should actually fall on a single line. However, measurement of the *y* variable is imprecise, which is seen as deviations from the line. Now, if the slope and intercept of the line were estimated from the set of points shown in Figure 4.2, it is likely that their values would not be exactly equal to the (unknown) true values, which describes the linear relationship between *x* and *y*. Furthermore, if the measurements of *y* were taken again, the measurement errors would generally take different values, and the slope and intercept estimates would change accordingly. Assume that this process was repeated *N* times, where *N* tends to infinity. If after *N* repetitions the slope and intercept estimates are averaged, and the average values are equal to the true slope and intercept values, the estimation method for the intercept and the slope is unbiased. Figure 4.3 shows the distribution of slope estimates after the *N* repetitions.

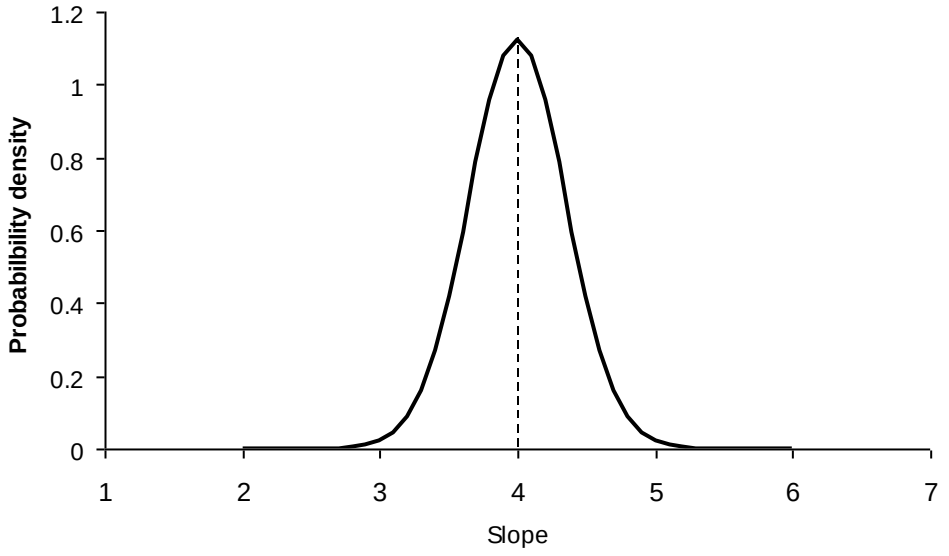

Figure 4.3. Distribution of occurrence of different slope estimates.

Although the slope clearly can take a range of values, the average over all repetitions is at 4. If we knew the true slope to be equal to 4, we would determine our estimate to be unbiased. In other words, bias in statistics is defined to be the difference between the true value and the expected value of the parameter being estimated. As the true parameter values are not normally known, conclusions about biasedness are usually based on assumed statistical properties of the distribution of error between the observed and modelled values, for instance that it has zero mean. Any systematic relation among the error values, such as correlation between successive ones, also generally affects statistical properties of the parameter estimates, such as their biases and variances.

EFFICIENCY

The most efficient estimate is defined to be the estimate with the minimum variance. Assume we have two methods to estimate the slope parameter in the example discussed above, and the two methods produce the distributions of occurrence shown in Figure 4.4. Both estimates are unbiased, but the estimate produced by the first method is more efficient than that produced by the second method, because it has a smaller variance (spread) around its mean value. Efficiency is a desirable property, as typically the observation data are observed only once and then, in the absence of bias, the more efficient the estimate is the more likely it is that the estimated slope is close to the true value.

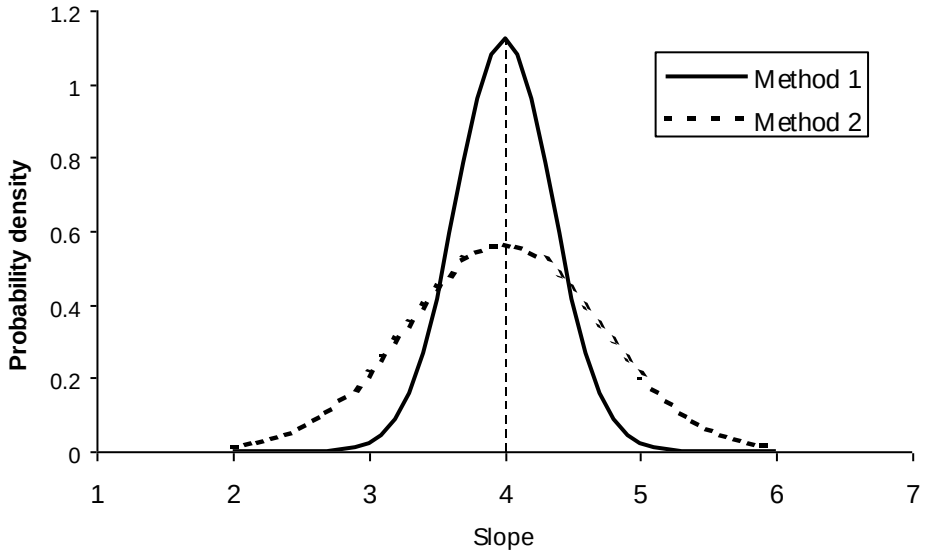

Figure 4.4. Distribution of occurrence of different slope estimates for two estimation methods.

MEAN SQUARE ERROR

While both unbiasedness and efficiency are desired properties for an estimate, it is really their combined effect that is important. Let us continue to consider the example of estimating the slope for the data shown in Figure 4.2a. Figure 4.5 presents two pairs of distributions for the slope. Each distribution is obtained using different estimation methods.

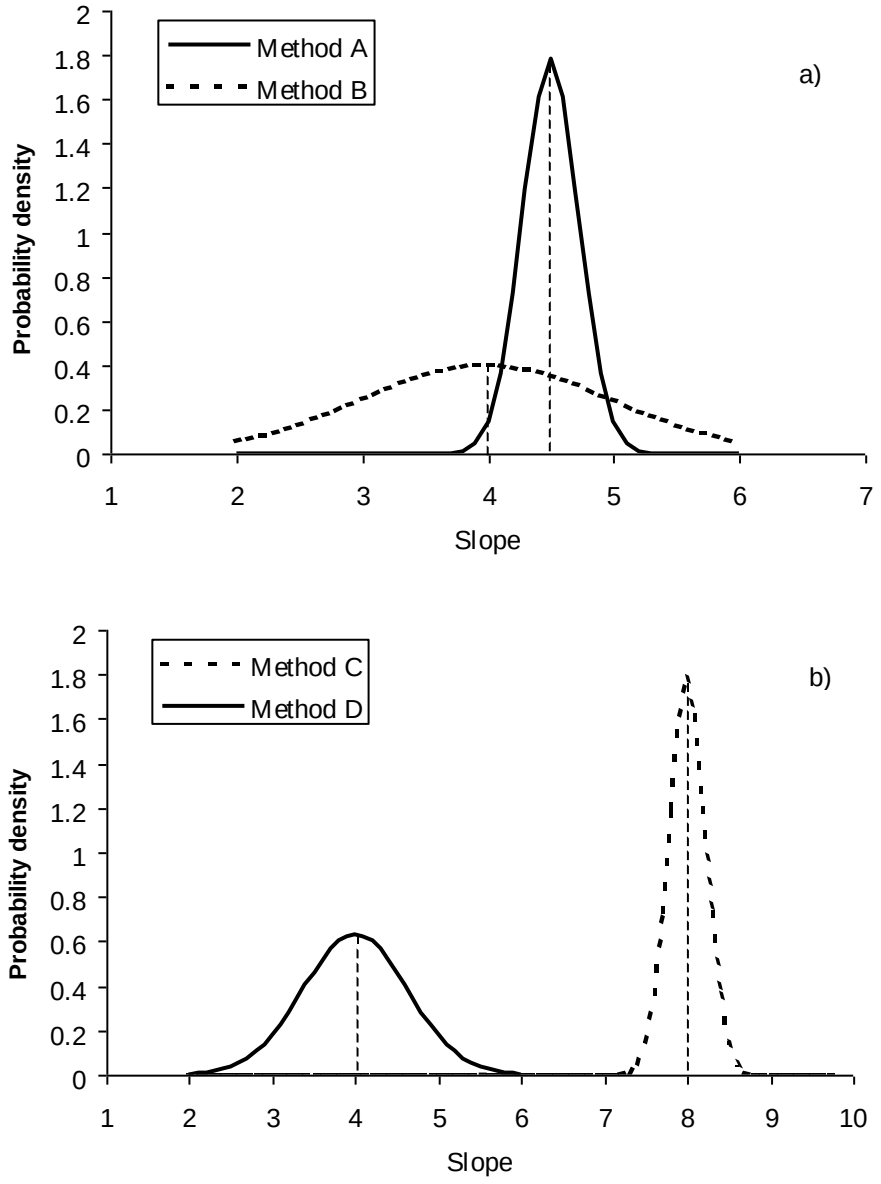

Figure 4.5. A biased, but efficient, estimation Method A vs. an unbiased, but inefficient, estimation Method B (a); and an efficient, but biased, estimation Method C vs. an inefficient, but unbiased, estimation Method D (b).

Clearly, in Figure 4.5a the Method A is the better of the two, although it is biased. Its better efficiency over Method B compensates for the slight bias of 0.5 (recall that the true slope value was assumed to be 4). And in Figure 4.5b, it is obvious the badly biased Method C, despite its greater efficiency, is inferior to Method D. Indeed, due to its efficiency (small variance) but large bias it is almost guaranteed to produce poor estimates for slope.

The mean square error $\tilde{x}$ takes into account both bias and efficiency, and therefore gives a better indication of the performance of an estimation method than any of the two alone. The mean square error is equal to the sum of the bias squared and the variance of an estimate, regardless of the error distribution.

CONSISTENCY

An estimate is said to be consistent when it tends to the true value as the number of observations tends to infinity. There are several mathematical definitions of consistency, depending on how convergence to the true value is defined (e.g. Papoulis, 1991).

CONFIDENCE INTERVALS

As its name suggests a confidence interval is the interval which contains the true value of a given parameter at the given level of confidence. For example, assume we estimated the slope of a line from the data shown in Figure 4.2a and obtained 3.8, for instance, as the slope estimate. Knowing that there is error in the measurements it seems reasonable that the true slope value is not necessarily exactly the estimated value. How could one assess what is the range in which the true slope value lies with a given probability?

Knowing the probability distribution of a random variable allows us to compute the interval where the value of the random variable lies with a given probability. Figure 4.6 shows the probability density function of the normal distribution with mean of 0 and variance of 64. The shaded area in the figure depicts 95% of the total area under the probability density curve. Now when the mean and the variance of the normal distribution are known, the 95% confidence interval around the mean of the distribution can be constructed. In this case the 95% confidence interval is from –15.7 to 15.7, i.e. the value of the random variable lies with probability 0.95 in the above range.

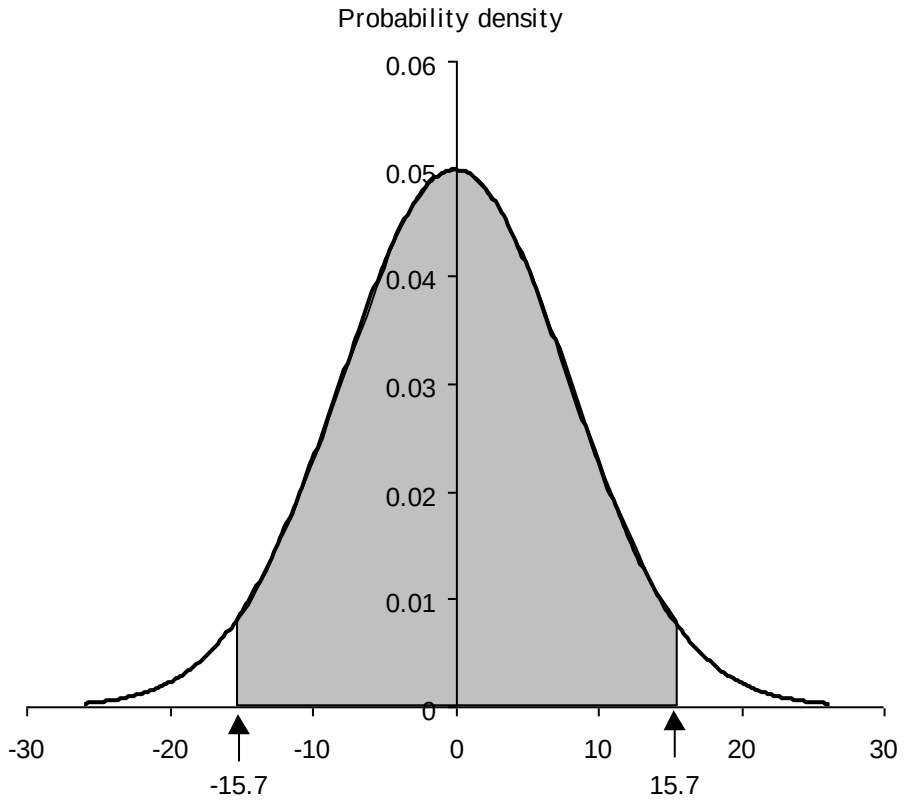

Figure 4.6. Probability density function of the normal distribution having a mean of 0 and variance of 64, and the associated 95% confidence interval.

As the 95% confidence interval contains by definition 95% of the probability mass and the normal distribution density function is symmetrical about the mean, the tails at both ends contain 2.5% of the probability mass each. Now the upper limit of the 95% confidence interval is at the 0.975 percentile, i.e. at 15.7 in the current example. Because of the symmetry of the normal distribution the lower limit resides at the negative value of the 0.975 percentile (obviously -15.7 in the present example).

In practical problems where the random variable is assumed to follow the normal distribution the variance of the distribution is rarely (if ever) known a priori and hence it needs to be estimated from the data. Then instead of using the normal distribution for locating the value at the desired percentile one should resort to Student's t-distribution, which takes into account uncertainty resulting from estimating the variance from a sample. Of course with an increasing sample size the uncertainty involved in estimating the variance decreases and Student's t-distribution (e.g. http://en.wikipedia.org/wiki/Student%27s_t-distribution) tends to the normal distribution as the sample size tends to infinity.

4.2 Ordinary least squares

THEORY

Let us start with a very simple linear model: a line characterising how variable *y* depends on variable *x*. Figure 4.7a depicts two points connected with a line. Now the two points are the data, the equation giving the line is the model (structure), and the intercept *a* and the slope *b* are the model parameters. As two points are sufficient to define a line fully, values for the two parameters – the intercept and the slope – can be found analytically. If a third data point became available and it fell on the line, no extra information on the values of the parameters would be obtained, but our confidence in the line being a good model structure would increase. If the new data point did not fall on the line, we would conclude that either data have been erroneously measured or the line is not the correct model structure for describing the process exactly.

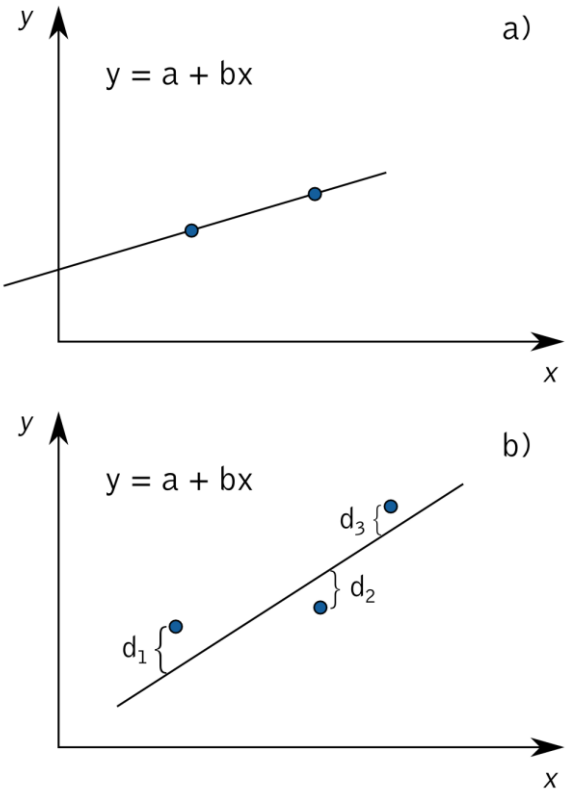

Figure 4.7. Schematic of a line.

In real life we have to tolerate errors arising from both inaccurate measurements (*measurement error*) and the fact that a model, being a simplification of Nature, cannot fully represent the real process with any parameter values (*model structure error*). In the light of these unavoidable errors, how should we draw a line through the set of three points shown in Figure 4.7b?

A reasonable way to do it is to define a metric or objective function that describes how well the line fits the data. A commonly used objective function is the least-squares criterion, i.e.

$$SSE = \sum_{n=1}^{N} (\hat{y}_n - y_n)^2 = \sum_{n=1}^{N} e_n^2 \tag{4.7}$$

where *SSE* is the sum of squared errors, $\hat{y}_n$ is the modelled value, y_n is the measured value, e_i is the discrepancy (error) between the modelled and measured values, and N is the number of data points. The least-squares method is attractive as it automatically penalises large errors much more than small errors, which is often (not always) desirable, and as it has some good statistical properties under suitable assumptions[5]. Now, the least-squares optimal line minimises the sum of squared vertical deviations (d_1, d_2, d_3 in Figure 4.7b) to the measured data, i.e.

$$\left[a_{opt}, b_{opt}\right] = \arg\min\left[d_1^2(a,b) + d_2^2(a,b) + d_3^2(a,b)\right] \tag{4.8}$$

where a_{opt} and b_{opt} are the optimal intercept and slope values, respectively.

We know from elementary calculus that for a smooth (local) minimum it is required that the first derivative be zero and the second derivative positive. Let us now write the problem of finding the optimal intercept and slope in a matrix form, which is convenient for presenting the analysis and will also cover more elaborate models. A good, concise summary about basics of matrix calculus can be found at http://en.wikipedia.org/wiki/Matrix_(mathematics). We can write for the case shown in Figure 4.7b

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} \tag{4.9}$$

If we now denote the observation vector as **y**, the matrix of explainers as **X**, the parameter vector (*a* and *b*) as **p**, and the deviation (error) vector as **e**, we have

$$\mathbf{y} = \mathbf{X}\mathbf{p} + \mathbf{e} \tag{4.10}$$

From (4.10) it follows that the error vector **e** can be written as

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{p} \tag{4.11}$$

The sum of squared errors can now be phrased using (4.11) as

$$\begin{aligned} SSE = d_1^2 + d_2^2 + d_3^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e} &= \left(\mathbf{y} - \mathbf{X}\mathbf{p}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{p}\right) \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{p} + \mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} \end{aligned} \tag{4.12}$$

Note that terms $\mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{p}$ are both scalars and identical, from which follows the equality between the second and third line in (4.12). Requiring the first derivative of *SSE* with respect to the parameter vector **p** to be zero yields

$$\frac{\partial SSE}{\partial \mathbf{p}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} = 0 \tag{4.13}$$

Solving for the parameter vector **p** from (4.13) gives the least squares estimate $\mathbf{p}_{LS}$ as

[5] The desired statistical properties and the required assumptions are listed in beginning of Section 4.3.

$$\mathbf{p}_{LS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{4.14}$$

To check that the stationary value of *SSE* given by $\mathbf{p}_{LS}$ is a minimum, notice that the Taylor series gives the change in *SSE* due to any small change $\Delta\mathbf{p}$ about $\mathbf{p}_{LS}$ as $\left(\mathbf{X}\Delta\mathbf{p}\right)^T\mathbf{X}\Delta\mathbf{p}$, which as a square form is positive[6].

Equation (4.14) is a powerful tool for parameter estimation, and it applies as such also to the multilinear model where there are more than one explanatory variables explaining the dependent variable *y*. Each additional explanatory variable just adds one column to the matrix **X**, and one element to the parameter vector **p**.

As we discussed earlier, the model result will differ from the data due to inaccuracies both in the measurements and in the model structure. These inaccuracies are reflected as uncertainty in the parameter vector **p**.

For the sake of clarity, assume for a moment that the linear model structure is fully correct, so all deviations between our fitted model and the observations arise from measurement errors. Of course we do not know how large those errors are for each individual observation. If we did, we could easily calculate the correct value and hence obtain an error-free observation record. However, we might have an idea what the magnitude of the squared error is on average, which allows us to consider the error process to be a random process having a zero mean and a known variance σ^2. If we further assume that the error follows the normal distribution, we can write for the error vector **e**

$$\mathbf{e} = \begin{bmatrix} \mathrm{Rnd}\left[N(0,\sigma^2)\right] \\ \mathrm{Rnd}\left[N(0,\sigma^2)\right] \\ \vdots \\ \mathrm{Rnd}\left[N(0,\sigma^2)\right] \end{bmatrix}$$

where Rnd[$N(\mu, \sigma^2)$] is a random number drawn from the normal distribution having a mean of μ and a variance of σ^2. Now, as the error is a random number it can take different values even for the same observation when the measurement is repeated. If the measurement campaign, which produced the three observations depicted in Figure 4.7b, were run again the deviations from d_1 to d_3 (and as a consequence the observations from y_1 to y_3) would generally take different values. As a result the least-squares estimates of the intercept and the slope would change. Obtaining different parameter estimates for each measurement campaign, which is due to the uncertainty contained in the observations, should make us ask how much confidence we can have in the derived parameter estimates.

As the parameter estimates depend on the measurements, and hence on the measurement errors, they are themselves random numbers. If the parameter estimate is unbiased, the mean of the parameter estimate equals the true parameter value. This means that if the measurement campaign were repeated many times, the average of the parameter values obtained from all those campaigns would approach the true parameter value. The covariance of the parameter estimate gives information about the accuracy of the parameter estimate derived from any single campaign. Covariance of a vector **v** is defined as

$$\mathrm{Cov}(\mathbf{v}) = E\left[(\mathbf{v}-\bar{\mathbf{v}})(\mathbf{v}-\bar{\mathbf{v}})^T\right] \tag{4.15}$$

where *E* is the expected value operator and $\bar{\mathbf{v}}$ is the mean of the vector **v**. The covariance matrix of the parameter estimates is thus a *P* x *P* matrix (*P* is the number of parameters) where variances of individual parameters are in the diagonal, and off-diagonal elements represent the covariances between the parameters.

Inserting the least squares parameter estimate $\mathbf{p}_{LS}$ as **v** in (4.15) yields

$$\mathrm{Cov}(\mathbf{p}_{LS}) = E\left[(\mathbf{p}_{LS}-\bar{\mathbf{p}}_{LS})(\mathbf{p}_{LS}-\bar{\mathbf{p}}_{LS})^T\right] \tag{4.16}$$

[6] See Appendix 4.1 on page 31 for derivation of this result.

Expressing $\mathbf{p}_{LS}$ as (see (4.14))

$$\mathbf{p}_{LS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\mathbf{p}+\mathbf{e}\right) = \mathbf{p} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{e} \tag{4.17}$$

recognising that $\bar{\mathbf{p}}_{LS} = \mathbf{p}$ (unbiasedness[7]), and defining $\mathbf{A} \equiv \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ allows us to write (4.16) as

$$\begin{aligned} \mathrm{Cov}(\mathbf{p}_{LS}) &= E\left[(\mathbf{Ae})(\mathbf{Ae})^T\right] = E\left[\mathbf{Aee}^T\mathbf{A}^T\right] = \mathbf{A}E\left[\mathbf{ee}^T\right]\mathbf{A}^T \\ &= \mathbf{A}(\sigma^2\mathbf{I})\mathbf{A}^T = \sigma^2\mathbf{AA}^T \end{aligned} \tag{4.18}$$

Now

$$\begin{aligned} \mathbf{AA}^T &= \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]^T = \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]\left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \tag{4.19}$$

so from (4.18)

$$\mathrm{Cov}(\mathbf{p}_{LS}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{4.20}$$

Given the observational set-up contained in the **X** matrix, the covariance matrix indicates how the uncertainty in the observations is translated into uncertainty in the estimated parameter values. The smaller the variances related to individual parameters are for a given data set, the more efficient is the estimate. It is important to note that even if we had no control over the magnitude of the measurement error (variance of **e**), we can decrease the uncertainty in the parameter values by designing wisely the observational set-up contained in **X**. The accuracy of the parameter estimates can be enhanced by increasing the number of observations (rows in **X**), but the problem of experimental design has more to it than just the sheer number of observations. Interested readers are referred, for example, to Montgomery (2008) or Ryan (2007).

Knowing the covariance structure of the model parameters, and assuming the model error to be normally distributed, allows us to construct confidence intervals for the model parameters.

EXERCISE 4.1

Objective

The objective of this exercise is to learn what a linear regression model is, and how its parameters can be estimated using the sum of squared errors criterion.

Background

You have a data set from 25 catchments, and your task is to identify which factors could be exploited to predict the spring time maximum daily runoff arising from snow melt. The available variables are:

- spring time daily maximum runoff q_{max} [mm/d]
- maximum snow water equivalent from the previous winter SWE_{max} [mm]
- lake percentage of a catchment *LP* [%]
- agricultural land percentage of a catchment *AP* [%]

You have all the data in an Excel workbook called CMWRA_4_1.xls (sheet Ex4.1a).

[7] See beginning of Section 4.3 for requirements of unbiasedness.

Task

- **Exercise 4.1a**

Construct a linear model, where the spring-time daily maximum runoff is explained by all the remaining variables. Note that the cells containing parameter values have been given names (par0 – par3). You can use these names when typing Excel worksheet functions.

Solve the parameter values using the sum of squared errors as the objective criterion. Do this with the Solver tool included in Excel (see Section 2.1). Have a look at the graph showing the predicted and measured maximum runoffs, and answer the following questions:

1. Are you satisfied with the model performance?
2. Do you find anything strange in the predicted values of maximum runoff?

- **Exercise 4.1b**

Now estimate the parameter values of the linear model yourself – without using the in-built Solver tool. Do this with matrix algebra (see (4.14)) using matrix operations included in Excel. You will need the following matrix formulae

- MMULT (multiplication of matrices)
- MINVERSE (inverse of a matrix)
- TRANSPOSE (transpose of a matrix)

In the CMWRA_4_1.xls workbook you will find a template for this exercise in sheet Ex4.1bc. Note that many matrices have been given names, which you may find helpful when typing your matrix functions. Answer the following question:

1. Do you obtain the same parameter values as when you minimised the sum of squared errors with Solver in Exercise 4.1a?

EXERCISE 4.2

Objective

The objective of this exercise is to understand the concepts of bias, efficiency and consistency, and to learn what factors affect the reliability of the estimated parameter values.

Background

Open the workbook CMWRA_4_2.xls and have a look at its contents. In essence it is a tool for generating pairs of data with a linear relationship between them, and for estimating parameter values (intercept and slope) of the line that the data pairs form. You can have some control on how the data pairs are generated. Let us start from the upper left corner. In the box titled *Observ.* you can change the number of observations you have and the range of the explanatory variable x_1. **Note: After having edited the values in this box you need to click the *Change* button!** Try it once or twice and observe what happens. In the next box, titled *True Process*, you specify the parameters of the line. The *Error Process* box gives you the possibility to control the uncertainty of both the explanatory variable x_1 and the dependent variable y.

The *Least Squares Est.* box shows the least squares estimates of the parameters for the current data. You do not need to do anything here – you know already how they are estimated. And in the last box – *Realisations* – you can control how many realisations of data you have available for the analysis.

The button *Get them!* will give you the specified number of realisations, and the button *Reset* will reset the total number of realisations back to zero. **Note: All the cells that are meant to be changed/edited by hand are in blue background and have a white font!**

Task

- **Exercise 4.2a**

Calculate – using *Excel* matrix functions – an estimate of the covariance of the linear model parameters in the box called *Theoretical Covariance* (see (4.20)). Explore how

1. errors in the dependent variable
2. the number of observations
3. the range of the explanatory variable x_1

affect the accuracy of the parameter estimates. Report your findings and discuss them.

- **Exercise 4.2b**

Use the following control variables:

- Number of observations: 21
- Range of x_1: 10
- Intercept: 5
- Slope: 2
- σ_y: 2
- σ_x: 0
- Number of realisations: 1

Click on the *Get them!* button for some twenty times and examine the behaviour of the graph for the fitted line in the chart in the bottom right corner. When you look at the estimated parameter values for the different realisations you observe that they are not the same.

1. Explain why the estimated parameter values for the different realisations are not the same.

- **Exercise 4.2c**

Use the *Realisations* tool in this workbook to explore how error in the dependent variable *y*, or in the explanatory variable x_1, affect the bias of the least squares parameter estimates. Answer the following questions:

1. Explain on what grounds you conclude that the parameter estimates are biased or unbiased, when there is error in the dependent variable *y*.
2. Explain on what grounds you conclude that the parameter estimates are biased or unbiased, when there is error in the explanatory variable x_1.
3. Investigate also how error in the dependent variable *y* affects the efficiency of the parameter estimates. Report your finding.
4. Are the parameter estimates consistent when there is error in the dependent variable *y*? Explain how you conclude this.
5. Are the parameter estimates consistent when there is error in the explanatory variable x_1? Explain how you conclude this.
6. What can you say about bias of the parameter covariance estimates when there is error in the dependent variable *y*?

4.3 Generalised least squares

THEORY

It is known that the ordinary least squares (OLS) estimate (4.14) is the most efficient, unbiased, linear estimate under the following assumptions:

1) The relationship between **x** and **y** is linear, i.e. in the case of two variables the straight line is the correct model structure
2) There is no error in the **x** values
3) The measurement error in **y** has zero expected value and a constant variance, and there is no correlation between errors

In such a case the covariance matrix of the error is

$$\mathrm{cov}(\mathbf{e}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I} \tag{4.21}$$

If we transformed the original data so that the covariance matrix of the transformed errors took this form, we would know that the least squares parameter estimates derived using the transformed data was the most efficient one. In the general case the covariance matrix of the error is

$$\mathrm{cov}(\mathbf{e}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2N} \\ \vdots & & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix} = \mathbf{C} \tag{4.22}$$

where **C** is a positive definite (or positive semi-definite[8]) matrix. A basic theorem of matrix algebra guarantees that there exists an invertible (non-singular) matrix **H** such that for positive-definite **C**

$$\mathbf{H}\mathbf{C}\mathbf{H}^{\mathsf{T}} = \mathbf{I} \tag{4.23}$$

From (4.23) it follows that

$$\mathbf{C} = \mathbf{H}^{-1}(\mathbf{H}^{\mathsf{T}})^{-1} = (\mathbf{H}^{\mathsf{T}}\mathbf{H})^{-1} \tag{4.24}$$

Next we will use the matrix **H** to transform the original data of (4.10), and obtain

$$\mathbf{H}\mathbf{y} = \mathbf{H}\mathbf{X}\mathbf{p} + \mathbf{H}\mathbf{e} \tag{4.25}$$

or

$$\mathbf{y}_{\mathbf{H}} = \mathbf{X}_{\mathbf{H}}\mathbf{p} + \mathbf{e}_{\mathbf{H}} \tag{4.26}$$

where $\mathbf{y}_{\mathbf{H}} = \mathbf{H}\mathbf{y}$, $\mathbf{X}_{\mathbf{H}} = \mathbf{H}\mathbf{X}$, and $\mathbf{e}_{\mathbf{H}} = \mathbf{H}\mathbf{e}$. Now the covariance matrix of the transformed error becomes

$$\mathrm{cov}(\mathbf{e}_{\mathbf{H}}) = E(\mathbf{e}_{\mathbf{H}}\mathbf{e}_{\mathbf{H}}^{\mathsf{T}}) = E(\mathbf{H}\mathbf{e}\mathbf{e}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}) = \mathbf{H}\sigma^2\mathbf{C}\mathbf{H}^{\mathsf{T}} = \sigma^2\mathbf{H}\mathbf{C}\mathbf{H}^{\mathsf{T}} = \sigma^2\mathbf{I} \tag{4.27}$$

From (4.27) we see that the transformed error is homoscedastic (i.e. it has constant variance) and has no correlation between its elements. Hence, when the estimate for parameter **p** is derived from (4.26), we know it to be unbiased and to have the minimum variance. We have

$$\mathbf{p}_{\mathrm{GLS}} = \left(\mathbf{X}_{\mathrm{H}}{}^{\mathsf{T}}\mathbf{X}_{\mathrm{H}}\right)^{-1}\mathbf{X}_{\mathrm{H}}{}^{\mathsf{T}}\mathbf{y}_{\mathrm{H}} \tag{4.28}$$

[8] If at least one error is always zero the **C** matrix is only positive semi-definite.

where $\mathbf{p}_{GLS}$ is the generalised least squares estimate for **p**. The generalised least squares (GLS) estimate yields the most efficient, unbiased linear estimate when the error term is heteroscedastic (i.e. variance of the error is not constant) and it is correlated. The GLS estimate coincides with the OLS estimate, as it should, when the error term is homoscedastic and non-correlated.

In order to apply (4.28), we need the matrix **H**, which again is derived from the covariance matrix of error **e**. As this covariance information is not usually known in advance, it needs to be estimated from the data. And as there are $N(N+1) / 2$ different elements in the covariance matrix, which all can have different values, it is clear that some more information needs to be available before those covariance values can be estimated. For example, if we know that the errors are not likely to be mutually correlated, but the error variance for the second half of observations is larger than for the first half of observations, we will have two different values in the principal diagonal of the covariance matrix. Now the generalised least squares estimation would proceed as follows:

1) Derive an OLS estimate
2) Estimate the two variances from the error series
3) Form the **H** matrix from the estimated covariance matrix
4) Transform the data using **H**
5) Derive an OLS estimate for the transformed data
6) Go to step 2

The estimation procedure would repeat steps from 2 to 6 until the parameter estimate did not significantly change.

EXERCISE 4.3

Objective

The objective of this exercise is to give an introduction to generalised least squares estimation, i.e. what can be achieved when the error term is heteroscedastic or correlated.

Background

Generalised least squares estimation is nothing else but derivation of an ordinary least squares estimate for a transformed data set. Transformation is based on the (known or estimated) covariance matrix of the error term.

In the workbook CMWRA_4_3.xls you will find a template for this exercise. The Ex4.3_Main sheet is very similar to the Ex4.2 sheet of the previous exercise. Here the number of observations and the range of the explanatory variable, however, cannot be changed. Again, all the values that are meant to be changed by hand are in white font and blue background.

To be able to compare OLS and GLS estimates, you have estimated parameter values and sample statistics for both these methods. Note that in the box *GLS Estimate* the messy looking formula is nothing else but the by now familiar (4.14), written in one line for the transformed data. Note also that you have under *Original data* an entry called *e Non-White*. This is the error that has a structure that is consistent with the error covariance matrix.

In addition, there is a sheet called Ex4.3_ErrorTerm. This is where the transformation matrix **H** of (4.25) is constructed based on the covariance matrix of the error term.

Task

- **Exercise 4.3a**

Begin by constructing the transformation matrix **H**. In this exercise it is assumed that you know the covariance structure of the error term. Let us further assume that the error term is not cross-correlated (all off-diagonal elements in the error covariance matrix are zeros), but it is heteroscedastic in such a way that the variance is different for the first 25 and for the last 25 errors. You can change these variance values in the *Error Process* box.

The transformation matrix **H** is required to satisfy (4.23), and it can be derived as the inverse of the so-called Cholesky decomposition of the covariance matrix. In case you are interested in knowing more about the Cholesky decomposition, see e.g. http://en.wikipedia.org/wiki/Cholesky_decomposition.

Below the covariance matrix you will find a place for the matrix that results when the covariance matrix is decomposed. Create this matrix with the custom worksheet matrix function CHOL(*matrix*) that is supplied in this workbook. Give as an argument – *matrix* – name or range of the matrix to be decomposed. After this, take the inverse of the Cholesky decomposed covariance matrix. You will find space for this matrix further down below. Now you have the transformation matrix **H** (the cell range containing it has the name H that you can use in worksheet functions later on) and can proceed to generalised least squares parameter estimation.

- **Exercise 4.3b**

Use the newly created transformation matrix **H** to transform the original data in the Ex4.3_Main sheet. Then use the *Realisations* tool to demonstrate that

1. the ordinary least squares estimate is unbiased, although the error term is heteroscedastic
2. the ordinary least squares estimate is not the most efficient of all linear estimates when the error term is heteroscedastic

Please report how you argue the above points based on the results you obtain using the *Realisations* tool.

EXERCISE 4.1 BRIEFLY REVISITED

- **Exercise 4.1c**

Now you know how to derive a covariance matrix for the parameter estimates of a linear model. This enables you to construct confidence intervals for the parameters. Construct 95% confidence intervals for the linear model parameters of Exercise 4.1. You can recall what a confidence interval is from Section 4.1. Note also that the variance for the error between observed and modelled values should be estimated from

$$s^2 = \frac{1}{N-M-1}\sum_{n=1}^{N} e_n^2 \tag{4.29}$$

where s^2 is the estimated variance, N is the number of observations, M is the number of explanatory variables, and e_n is the discrepancy (error) between the n^{th} modelled and the n^{th} measured value. The reason for the divisor in the above equation being $N - M - 1$ is that, while there are N data points, there are only $N - M - 1$ independent observations as the $M + 1$ regression parameters put $M + 1$ constraints on the data. The number of independent observations is referred to as the number of degrees of freedom. Interested readers can find more information at http://en.wikipedia.org/wiki/Degrees_of_freedom_%28statistics%29.

You have a template for this in CMWRA_4_1.xls workbook in sheet Ex4.1bc. Answer the following questions:

1. Which factors explain the spring-time maximum flow at the 95% level of confidence?
2. How do you conclude this from the confidence intervals?

4.4 Non-linear optimisation

THEORY

Often the process to be described cannot properly be represented with a (multi)linear model, and a non-linear model structure is required. To construct a method for estimating parameter values for a non-linear model, we start again from the requirement that the first derivative of the objective function with respect to the parameter vector be zero. Imagine that we are adjusting the parameter vector **p** in steps and have reached a value $\mathbf{p}_k$ after step k. Now, with aid of the Taylor's expansion we can approximate the first derivative of the objective function J close to $\mathbf{p}_k$ as

$$\frac{\partial J}{\partial \mathbf{p}} \cong \frac{\partial J}{\partial \mathbf{p}_k} + \frac{\partial^2 J}{\partial \mathbf{p}_k^2} \Delta \mathbf{p}_k \tag{4.30}$$

where $\Delta\mathbf{p}_k$ is the displacement of the parameter value from $\mathbf{p}_k$. We can try to make the gradient $\frac{\partial J}{\partial \mathbf{p}}$ zero, by solving from (4.30) for $\Delta\mathbf{p}_k$

$$\Delta \mathbf{p}_k = -\frac{\frac{\partial J}{\partial \mathbf{p}_k}}{\frac{\partial^2 J}{\partial \mathbf{p}_k^2}} = -\mathbf{H}^{-1}\mathbf{g} \tag{4.31}$$

where the second derivative (Hessian matrix) is denoted by **H**, and the first derivative (gradient vector) by **g**. We will probably not achieve exactly zero first derivative of *J* by use of (4.31), because (4.30) is only an approximation that neglects all higher derivatives in the Taylor's expansion[9]. All we can do is use $\Delta\mathbf{p}_k$ from (4.31) to obtain a new parameter estimate $\mathbf{p}_{k+1} = \mathbf{p}_k + \Delta\mathbf{p}_k$. This is then inserted into (4.31), and the parameter displacement recalculated. This iterative procedure is continued until we are satisfied with the final parameter estimate. The estimation algorithm is said to converge monotonically (in *J*) if the next parameter estimate is always better, i.e. produces a smaller value of the objective function than the previous estimate. There are several possible stopping criteria for the iteration, a common one being that the parameter value does not significantly change from one iteration to the next, i.e. $\Delta\mathbf{p}_k$ is small.

The update formula (4.31) is called Newton's method and it dates back to the 17th century. As Newton's method requires calculation of the second derivative (the Hessian matrix), it belongs to the group of second-order methods. The Hessian matrix can be found analytically only in special cases, and the need to estimate it numerically causes Newton's method to bear a fairly heavy computational burden. Also, Newton's method tends to suffer from convergence problems when the initial parameter estimate is far from the optimum.

The Gauss-Newton method has been derived from Newton's method, and it is particularly suited for problems where the sum of squares – already known to us from the linear parameter estimation – is used as the objective function. Adopting (4.7) as *J* allows us to write

$$J = \sum_{n=1}^{N} \left(\hat{y}_n - y_n \right)^2 = \sum_{n=1}^{N} e_n^2 \tag{4.32}$$

The modelled value $\hat{y}_n$, and consequently the error e_n, are of course functions of the parameter vector **p**. The gradient vector **g** can now be expressed as

$$\mathbf{g} = \begin{bmatrix} 2\sum_{n=1}^{N} e_n \frac{\partial e_n}{\partial p_1} \\ 2\sum_{n=1}^{N} e_n \frac{\partial e_n}{\partial p_2} \\ \vdots \\ 2\sum_{n=1}^{N} e_n \frac{\partial e_n}{\partial p_P} \end{bmatrix} \tag{4.33}$$

where e_n is the model error with respect to the n^{th} observation, *N* is the number of observations, p_i is i^{th} parameter in the parameter vector **p**, and *P* is the number of parameters. The Hessian **H** can be written as

[9] When *J* has the form of sum of squared errors (see (3.1)) and the model is linear in **p**, the first order approximation is accurate and the problem simplifies to linear regression.

$$\mathbf{H} = \begin{bmatrix} 2\sum_{n=1}^{N}\left[\frac{\partial e_n}{\partial p_1}\frac{\partial e_n}{\partial p_1} + e_n\frac{\partial^2 e_n}{\partial p_1 \partial p_1}\right] & 2\sum_{n=1}^{N}\left[\frac{\partial e_n}{\partial p_1}\frac{\partial e_n}{\partial p_2} + e_n\frac{\partial^2 e_n}{\partial p_1 \partial p_2}\right] & \cdots & 2\sum_{n=1}^{N}\left[\frac{\partial e_n}{\partial p_1}\frac{\partial e_n}{\partial p_P} + e_n\frac{\partial^2 e_n}{\partial p_1 \partial p_P}\right] \\ 2\sum_{n=1}^{N}\left[\frac{\partial e_n}{\partial p_2}\frac{\partial e_n}{\partial p_1} + e_n\frac{\partial^2 e_n}{\partial p_2 \partial p_1}\right] & \ddots & & \\ \vdots & & \ddots & \\ 2\sum_{n=1}^{N}\left[\frac{\partial e_n}{\partial p_P}\frac{\partial e_n}{\partial p_1} + e_n\frac{\partial^2 e_n}{\partial p_P \partial p_1}\right] & \cdots & \cdots & 2\sum_{n=1}^{N}\left[\frac{\partial e_n}{\partial p_P}\frac{\partial e_n}{\partial p_P} + e_n\frac{\partial^2 e_n}{\partial p_P \partial p_P}\right] \end{bmatrix} \tag{4.34}$$

When the parameter **p** is not far from the optimal value, the error e_n is usually small enough to make the second-order terms in (4.34) negligible compared with the terms involving the product of first derivatives. This simplifies the expression for the Hessian into the form

$$\mathbf{H} \cong \begin{bmatrix} 2\sum_{n=1}^{N}\frac{\partial e_n}{\partial p_1}\frac{\partial e_n}{\partial p_1} & 2\sum_{n=1}^{N}\frac{\partial e_n}{\partial p_1}\frac{\partial e_n}{\partial p_2} & \cdots & 2\sum_{n=1}^{N}\frac{\partial e_n}{\partial p_1}\frac{\partial e_n}{\partial p_P} \\ 2\sum_{n=1}^{N}\frac{\partial e_n}{\partial p_2}\frac{\partial e_n}{\partial p_1} & \ddots & & \\ \vdots & & \ddots & \\ 2\sum_{n=1}^{N}\frac{\partial e_n}{\partial p_P}\frac{\partial e_n}{\partial p_1} & \cdots & \cdots & 2\sum_{n=1}^{N}\frac{\partial e_n}{\partial p_P}\frac{\partial e_n}{\partial p_P} \end{bmatrix} \tag{4.35}$$

As we will soon see, it is convenient to define now a new matrix **S** as

$$\mathbf{S} = \begin{bmatrix} \frac{\partial e_1}{\partial p_1} & \frac{\partial e_1}{\partial p_2} & \cdots & \frac{\partial e_1}{\partial p_P} \\ \frac{\partial e_2}{\partial p_1} & \ddots & & \\ \vdots & & \ddots & \\ \frac{\partial e_N}{\partial p_1} & \cdots & \cdots & \frac{\partial e_N}{\partial p_P} \end{bmatrix} \tag{4.36}$$

This matrix **S** is called the sensitivity matrix, which – as its name suggests – expresses the sensitivity of the model error to a change in the parameter value. Using the sensitivity matrix **S** we can rewrite the gradient **g** (see (4.33)) as

$$\mathbf{g} = 2\mathbf{S}^{\mathrm{T}}\mathbf{e} \tag{4.37}$$

and the Hessian as

$$\mathbf{H} \cong 2\mathbf{S}^{\mathrm{T}}\mathbf{S} \tag{4.38}$$

When we now insert (4.37) and (4.38) into Newton's method (4.31), we obtain

$$\Delta\mathbf{p} \cong -\mathbf{H}^{-1}\mathbf{g} \cong -\left(2\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1} 2\mathbf{S}^{\mathrm{T}}\mathbf{e} = -\left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{e} \tag{4.39}$$

Anything familiar in (4.39)? Indeed, it has exactly the same form as the equation for the linear least squares parameter estimation (4.14). But note also:

1) There is a negative sign on the right hand side

2) Instead of directly yielding an estimate for the parameter vector, the equation gives a displacement of the parameter vector, which is added to the parameter vector value from the previous iteration round
3) The **X** matrix accommodating the explanatory variables has been replaced with the sensitivity matrix **S**
4) The **y** vector accommodating the observed values has been replaced with the error vector **e**

Clearly the algorithm can be thought of as minimising the sum of squared errors in fitting **e** by $\mathbf{S}\Delta\mathbf{p}$.

It is noteworthy that even though the Gauss-Newton method derives from Newton's method, there is no longer any need to find second derivatives of J with respect to **p**. In forming the sensitivity matrix only first derivatives are required, which offers a substantial reduction in numerical effort.

The robustness of the Gauss-Newton method can be enhanced by taking special care in situations where the $\mathbf{S}^T\mathbf{S}$ matrix is (near-) singular and thus cannot be inverted. Inverting a singular matrix is equivalent to dividing by zero in the scalar case. Adding any positive numbers to the entire principal diagonal of $\mathbf{S}^T\mathbf{S}$ makes it non-singular and hence invertible. Therefore, the numerical robustness of the Gauss-Newton method can be improved by replacing the $\mathbf{S}^T\mathbf{S}$ matrix with $\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I}$, where λ is a positive coefficient and **I** is the unit matrix. Now selecting a positive value for λ guarantees $\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I}$ to be invertible. With this extension, the method is called the Levenberg-Marquardt method, and it can be written as

$$\Delta\mathbf{p} = -\left(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I}\right)^{-1}\mathbf{S}^T\mathbf{e} \tag{4.40}$$

We could, of course, use any reliable numerical optimisation algorithm to minimise J; the special point about the Gauss-Newton and Levenberg-Marquardt methods is that they exploit the sum-of-squares nature of the objective function.

4.5 Problems in parameter estimation

THEORY

Local optima

In parameter estimation it can easily be the case that the objective function we try to minimise has multiple (local) optima. Many parameter estimation methods, like the Gauss-Newton and Levenberg-Marquardt methods presented above, will only find a local minimum, which is not necessarily the global minimum. Figure 4.8 demonstrates the existence of multiple local optima for a model with a single parameter.

Now, if a parameter estimation method is prone to converge to the nearest local minimum, the initial value for starting the estimation process decides whether the optimal parameter value is identified to be p_1 or p_2. If p_1 is identified as the optimal parameter value, the global minimum at p_2 is not discovered.

The most straightforward method for locating the global minimum in the objective function is to exhaustively sample the parameter space and select then those parameter values that yield the lowest value for the objective function J. It may be feasible to sample the entire parameter space at sufficiently small intervals if the model only has few parameters, but when the number of parameters increases the computational burden can quickly become intractable. There are also more sophisticated methods for searching the global minimum. The interested reader is referred to e.g. Duan et al. (1992), http://en.wikipedia.org/wiki/Simulated_annealing, and Kirkpatrick et al. (1983).

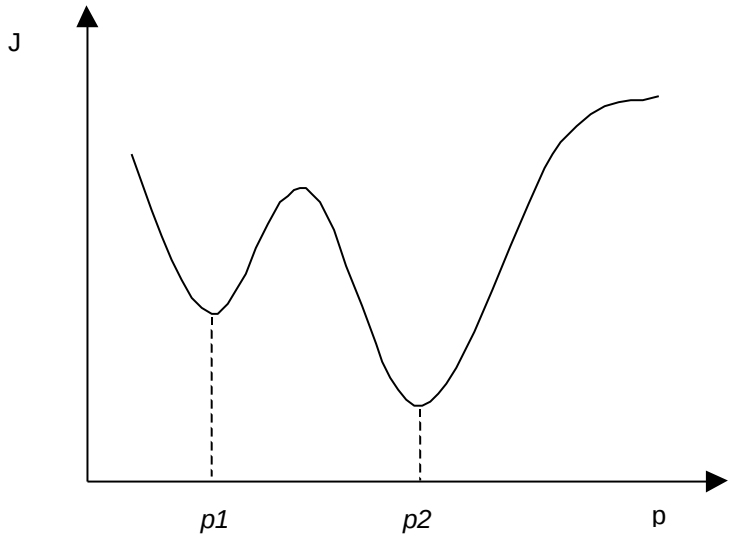

Figure 4.8. Existence of multiple local minima in the objective function *J* for a one parameter (*p*) model.

Non-uniqueness

If there is only one solution to a parameter estimation problem the problem is said to have a unique solution. In the opposite case many different combinations of parameter values result in the same (minimum) value of the objective function, and the problem is said to have a non-unique solution. Figure 4.9 depicts the situation where the solution is non-unique for a one parameter model.

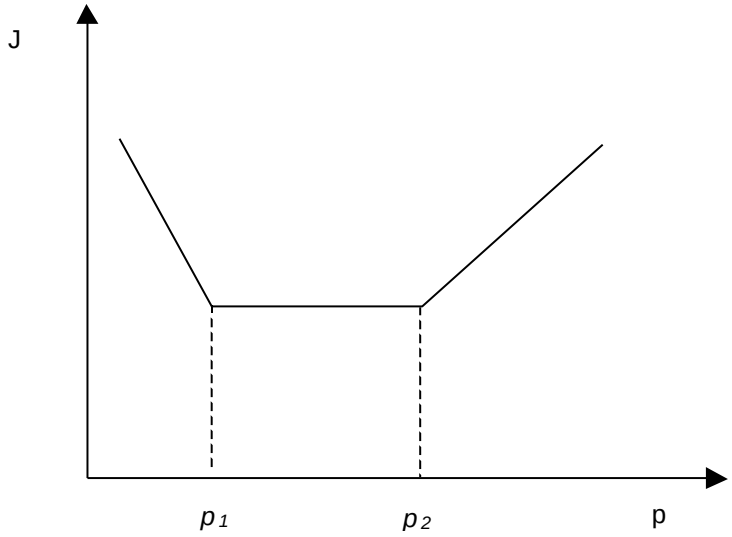

Figure 4.9. Existence of multiple local minima in the objective function *J* for a one parameter (*p*) model.

Clearly, any parameter value between p_1 and p_2 results in the minimum value of the objective function *J*. Why can this be a problem? Why should we be interested in the parameter values as long as we obtain the optimal value for the objective function? Non-uniqueness is not desirable as it suggests that the problem definition is not sufficient to identify the parameter values reliably. Although any of the several parameter value sets gives an optimal value of the objective function for that problem set-up that is used in parameter estimation, it is well possible that different parameter values yield different model outputs when the model is applied to a different case (for example, a different set of values of the independent variable).

A problem related to the non-uniqueness is the case where there may be a unique solution, but many optimisation methods have problems in finding it because the objective function surface is so flat. Figure 4.10 shows a case where many combinations of two parameter values give nearly an equally good value for the objective function.

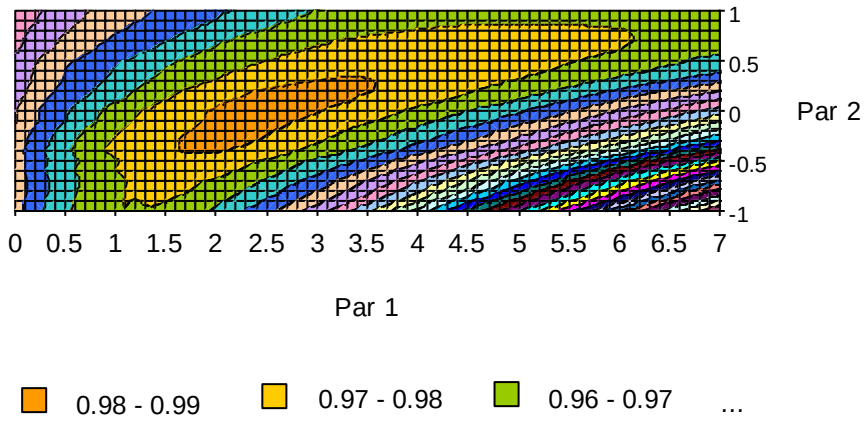

Figure 4.10. Contour plot of the objective function value for different combinations of parameters Par 1 and Par 2. The objective function is the Nash – Sutcliffe criterion whose maximum value is unity for the perfect match between measured and modelled values (see (7.9) in Section 7.1).

Again, although the model set-up used in parameter estimation could not properly differentiate between several parameter value combinations, it is quite possible that an application of the model to a different case shows significant differences in the model output for alternative parameter value combinations. For this reason, it is also quite possible that the parameter combination yielding the best fit in the calibration period does not give a good performance outside the calibration data, i.e. in validation (see Sections 3.1 and 3.2).

Instability

Even when there exists a clearly unique solution, it is a problem if the optimal parameter values do not smoothly depend on the objective function. Figure 4.11 demonstrates a problematic relationship between the value of the objective function J, and the optimal parameter value p_{opt}.

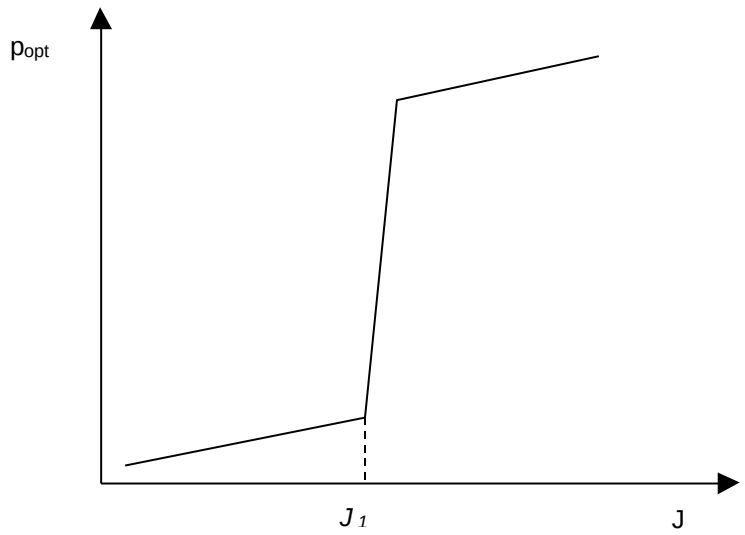

Figure 4.11. Relationship between the value of the objective function J, and the optimal parameter value p_{opt}.

It is easy to see that there is a jump in the dependence of the optimal parameter value p_{opt} on the objective function value at J_1. If the value of the objective function changes only slightly around J_1, the optimal parameter value changes dramatically. This is an undesirable property as even a small error in observations, which causes J to change slightly, can have a large impact on the optimised parameter values. Many parameter estimation problems in environmental studies, e.g. estimation of transmissivities from measured groundwater levels, are known to be prone to instability problems.

Regularization

All the difficulties listed above can be encapsulated in the concept of ill-posedness of the estimation problem. The ill-posedness can be alleviated by applying additional information to the estimation problem. For example, instead of using just the squared differences between measured and simulated values of the model output as an objective function (4.7), prior information on parameter values – when available – can be incorporated into the objective function. In such a case the objective function J can be written as

$$J(\mathbf{p}) = \sum_{n=1}^{N} [\hat{y}_n(\mathbf{p}) - y_n]^2 + \alpha^2 \sum_{i=1}^{M} (p_i - p_i^0)^2 \tag{4.41}$$

where p_i is the i^{th} parameter of the parameter vector, p_i^0 is the prior estimate for the i^{th} parameter of the parameter vector, and α is the regularization coefficient. Equation (4.41) defines the regularized estimation problem, where the value of α determines how much belief is put on the observed data, and on the initial parameter estimate. When α is zero the observed data alone determine the solution, whereas when α approaches infinity the solution becomes equal to the prior estimate.

The regularization method (Tikhonov, 1963) forces the objective function to take more of a quadratic shape, which improves the performance of many numerical estimation methods – including the Levenberg-Marquardt method presented above. The greater the value of α the more the objective function resembles the quadratic shape. Of course, it is often difficult to come up with a reasonable initial estimate of parameter values, and a large value of α implies that a lot of confidence is placed in the initial estimate. Choice of the value of α is a compromise between the performance of the estimation method, and the extent to which the observed data are exploited in identifying parameter values. Selection of the value of α should depend on the error between modelled and observed outputs. Interested readers are referred to Sun (1994) and references therein.

References

Kirkpatrick, S., Gelatt, C. D., and Vecchi, C. D. 1983. Optimization by simulated annealing, Science, 220(4598), 671-680.

Montgomery, D.C. 2008. Design and Analysis of Experiments (7th edition). Wiley, New York.

Ryan, T.P. 2007. Modern Experimental Design. Wiley.

Papoulis, A . 1991. Probability, Random Variables and Stochastic Processes (3rd edition). McGraw-Hill, New York.

Sun, N-Z. 1994. Inverse Problems in Groundwater Modeling. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Tikhonov, AN. 1963. Solution of incorrectly formulated problems and the regularization method. Soviet Math. Dokl. 4, 1035-1038, English translation of Dokl Akad. Nauk SSSR 151, 1963, 501-504.

Appendix 4.1 Change in SSE due to displacement of the parameter value from its ordinary least squares estimate

Taylor series gives the change in sum of squared errors (*SSE*) due to any small change $\Delta\mathbf{p}$ about $\mathbf{p}_{LS}$ as

$$(\Delta\mathbf{p})^T \frac{\partial SSE(\mathbf{p}_{LS})}{\partial \mathbf{p}} + \frac{1}{2!}(\Delta\mathbf{p})^T \frac{\partial^2 SSE(\mathbf{p}_{LS})}{\partial \mathbf{p}^2} \Delta\mathbf{p} \tag{4.42}$$

Now

$$\frac{\partial SSE}{\partial \mathbf{p}} = -2\mathbf{X}^{\mathsf{T}}\mathbf{y} + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{p}\,, \tag{4.43}$$

$$\frac{\partial^2 SSE}{\partial \mathbf{p}^2} = 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\text{, and} \tag{4.44}$$

$$\mathbf{p}_{\mathrm{LS}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{4.45}$$

Inserting first $\mathbf{p}_{\mathrm{LS}}$ from (4.45) for $\mathbf{p}$ in (4.43), and then (4.43) and (4.44) into (4.42) yields

$$\begin{aligned}
&\left(\Delta\mathbf{p}\right)^{T}\left[-2\mathbf{X}^{\mathsf{T}}\mathbf{y} + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}\right] + \frac{1}{2}\left(\Delta\mathbf{p}\right)^{T} 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\Delta\mathbf{p} \\
&= \left(\Delta\mathbf{p}\right)^{T}\left[-2\mathbf{X}^{\mathsf{T}}\mathbf{y} + 2\mathbf{X}^{\mathsf{T}}\mathbf{y}\right] + \left(\Delta\mathbf{p}\right)^{T}\mathbf{X}^{\mathsf{T}}\mathbf{X}\Delta\mathbf{p} \\
&= \left(\Delta\mathbf{p}\right)^{T}\mathbf{X}^{T}\mathbf{X}\Delta\mathbf{p} \\
&= \left(\mathbf{X}\Delta\mathbf{p}\right)^{T}\mathbf{X}\Delta\mathbf{p}
\end{aligned}$$

5 PRECIPITATION

T. Kokkonen

5.1 Areal precipitation

THEORY

Often in hydrological applications an estimate of average precipitation over a given area is needed. For example, an areal estimate is important in assessing water balance at the basin (catchment) scale.

There are several ways for deriving areal estimates from a set of point measurements. One of the classical approaches is the Thiessen polygon method (Thiessen, 1911) where the polygons bound regions that are closer to a given rainfall station relative to all other stations. In other words, the lines of the polygons are drawn so that they are of equal distance between two adjacent stations (Figure 5.1). The areal average estimate is then calculated as the weighted average of point rainfall values from all stations. The areas of the polygons – intersected with the region boundary – are used as weights (regions R1…R4 in Figure 5.1).

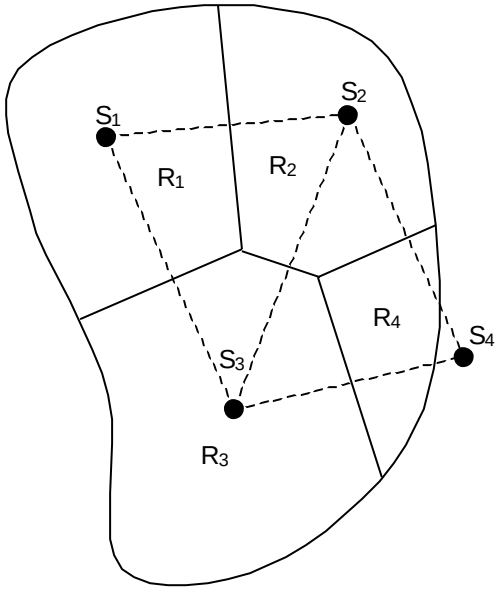

Figure 5.1. Thiessen polygons for estimating the areal average precipitation for the region shown in the figure. Rainfall stations are indicated by S_1…S_4, and regions are labelled with R_1…R_4.

The isohyetal method is based on drawing contours of equal precipitation (isohyets). The average precipitation of two adjacent isohyets is assumed to represent the precipitation amount in the region bounded by the two isohyets. The areal value is then derived analogously to the Thiessen method, i.e. it is a weighted average of precipitation values representing regions bounded always by two isohyets. The areas of the regions bounded by a pair of isohyets are used as weights. Using the notation of Figure 5.2, the areal precipitation P_A is computed from

$$P_A = \frac{\sum_{i=1}^{4} A_i 0.5\left(P_i + P_{i+1}\right)}{\sum_{i=1}^{4} A_i} \tag{5.1}$$

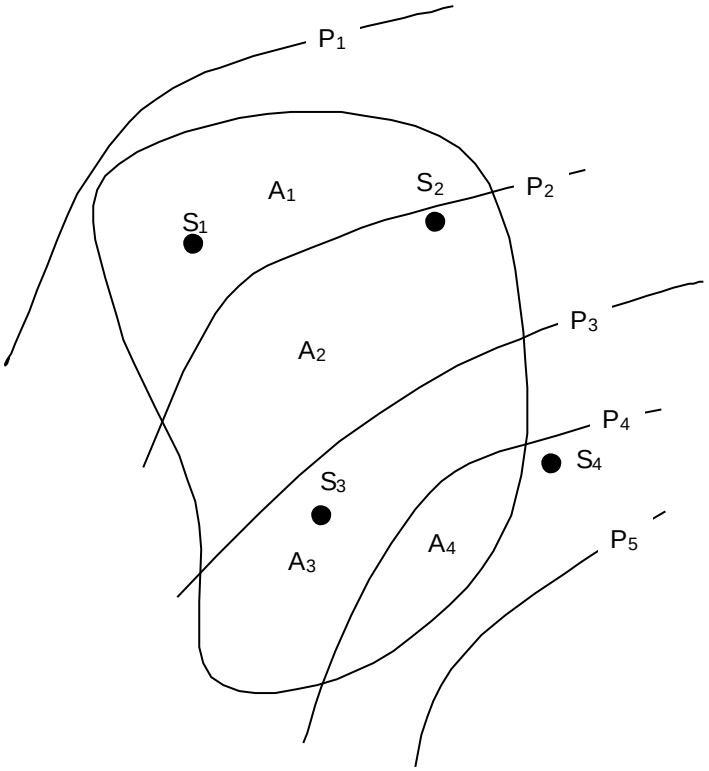

Figure 5.2. Isohyetal method for estimating the areal average precipitation for the region shown in the figure. Rainfall stations are indicated by $S_1 \ldots S_4$, areas of subregions are labelled with $A_1 \ldots A_4$, and isohyet values are labelled with $P_1 \ldots P_4$.

In the inverse distance method the region is subdivided into a regular grid, and the precipitation in each grid cell is taken as a weighted average of all stations considered. The weights are calculated as the powers of the inverse of the distance from the grid cell to a station. Typically the power has been assigned a value of 2 (Brutsaert, 2005). The spline method utilises spline functions to describe the relationship between point coordinates (latitude, longitude) and the precipitation value. The spline surface is first fitted to the available point precipitation data, and the fitted surface is used to derive an areal precipitation estimate. In kriging, which was originally developed for the mining industry for interpolating gold grades, the spatial correlation of the precipitation is exploited in constructing the areal estimate over a given region.

Precipitation typically has a relationship with elevation; the higher elevated regions tend to receive more precipitation. In the isohyetal method a person having experience in drawing isohyets can exploit knowledge of local topography to account for the effect of elevation. In other methods, such as spline fitting and kriging, elevation can also be taken into account in the derivation of areal averages.

References

Brutsaert, W. 2005. Hydrology – An Introduction. Cambridge University Press, New York.

Thiessen, A.H. 1911. Precipitation averages for large areas. Monthly Weather Review, 39, 1082-1084.

EXERCISE 5.1

Objective

The objective of this exercise is to learn how to derive an areal precipitation estimate from a set of point measurements using the Thiessen polygon method.

Meteorological data

The precipitation data available for this exercise were kindly provided by the Finnish Meteorological Institute. It is not permitted to use the data for purposes other than solving this exercise.

Task

This exercise is built around the Geoinformatica software package. See Section 2.2 for instructions on how to install and use Geoinformatica. The Vantaanjoki basin area and the rainfall station locations are available as two grid data sets called *vantaa_basin.tif* and *vantaa_stations.tif*, respectively. You also have the rainfall data for 10 rainfall stations in an Excel file called *vantaa_precipitation.xls*. After you have opened the grid data sets remember to load them to RAM (see the important note in *Opening data* in Section 2.2).

Use the Thiessen polygon method and compute areal precipitation for Vantaanjoki basin for two summers: Jun – Aug 2004, and Jun – Aug 2006. Answer the following questions:

1. What is the difference in the accumulated areal summer precipitation between the years 2004 and 2006?
2. Which rainfall station would have the largest – and which the smallest – impact on the estimated basin average rainfall if it ceased to operate?

Hints

You will find interpolation and zonal functions of Geoinformatica useful in this exercise. See Perl modules documentation (accessible through the Geoinformatica start menu) under *Classes-> Geo::Raster::Algorithms* and *Classes-> Geo::Raster*. Using the search tool of your html browser may help you to find those functions that you need more quickly.

You may also want to refer back to Section 2.2 about how to apply different colour schemes and how to arrange the map view in Geoinformatica.

6 EVAPORATION

L. Stenberg, T. Kokkonen

6.1 Lake evaporation

THEORY

Combination equation

Evaporation is the process whereby water in its liquid or solid state is converted to water vapour. This process is controlled by two factors: 1) the available energy at the evaporating surface, and 2) the humidity of the air above the evaporating surface. In transition from liquid water to water vapour the distance between water molecules increases greatly, and energy is needed to overcome the attractive intermolecular forces. When evaporation continues the air just above the evaporating surface becomes saturated, and evaporation ceases, unless some mechanism (such as wind) carries away the water molecules as they leave the surface.

Based on the two controlling factors, methods describing evaporation can be classified as mass transfer and energy balance formulations. It is worth noting, however, that this division is somewhat artificial as mass and energy balances are closely coupled in the evaporation process. Amount of evaporation can be considered as the depth of water that is transferred to water vapour, but just as well it can be considered as the amount of energy required to convert water from the liquid phase to the vapour phase.

The basis for mass transfer formulations is Dalton's law that states the rate of evaporation E be proportional to the water vapour pressure difference e_s-e_a

$$E \propto e_s - e_a \tag{6.1}$$

where E [kg m^{-2} d^{-1}] is the evaporation rate as a mass flux, e_s [kPa] is the vapour pressure at the water surface, and e_a [kPa] is the vapour pressure of the overlying air. The coefficient of proportionality is a function of wind speed, which arises from the need to carry moist air away from the evaporating surface (see above). Clearly for evaporation to occur the vapour pressure in the overlying air needs to be smaller than at the water surface. In the opposite case condensation prevails at the water surface and water vapour is converted to liquid or solid state.

Solar radiation emitted by the sun is a major component of the energy balance. Solar radiation is often referred to as short-wave radiation, which is in contrast with the radiant energy at longer wavelengths emitted by the earth and atmosphere. The characteristic wavelength of radiation emitted by an object is related to its surface temperature: a hot object emits radiation at shorter wavelengths than a cooler object. Part of the extraterrestrial short-wave radiation is absorbed in the atmosphere, and part of the short-wave radiation reaching the earth's surface is reflected. Net short-wave (long-wave) radiation comprises the sum of downward and upward components of short-wave (long-wave) radiation at the earth's surface (Figure 6.1). Net radiation is the net input of radiation at the surface, i.e. the sum of net short-wave and net long-wave radiation.

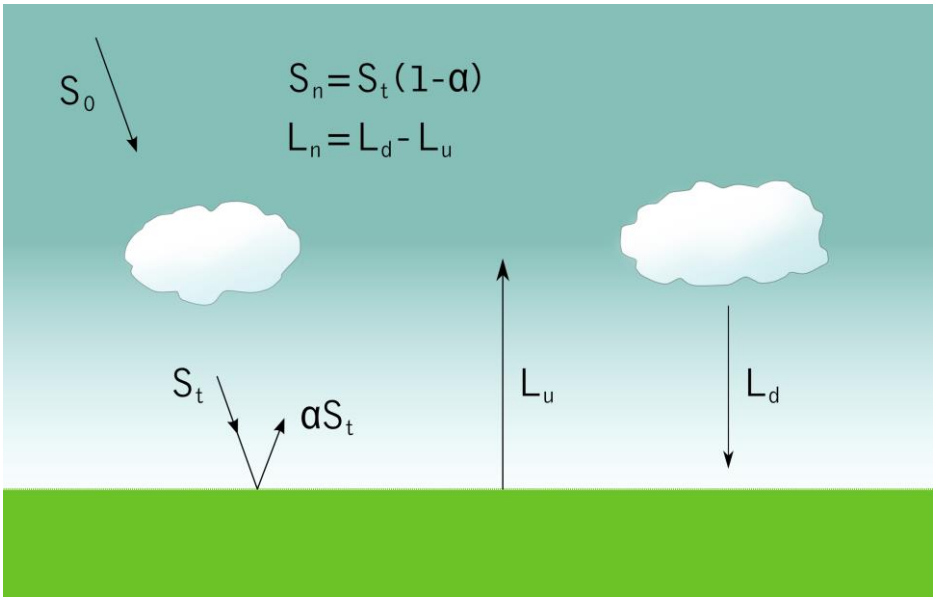

Figure 6.1. Radiation balance at the earth's surface. S_0 is the extraterrestrial short-wave radiation, S_t is that part of short-wave radiation reaching the earth's surface, αS_t (where α is the albedo) is the reflected short-wave radiation, L_u is the upward long-wave radiation emitted by the earth's surface, L_d is the downward long-wave radiation emitted by the atmosphere, S_n is the net short-wave radiation, and L_n is the net long-wave radiation. Note that the positive direction is towards the earth's surface.

Energy balance at the evaporating surface can be written as

$$R_n = \lambda E + H + G \tag{6.2}$$

where R_n [MJ m^{-2}d^{-1}] is the net radiation, λE [MJ m^{-2}d^{-1}] is the latent heat flux, H [MJ m^{-2}d^{-1}] is the sensible heat flux, and G [MJ m^{-2}d^{-1}] is the heat conduction into the ground or water beneath the evaporating surface. Here sensible heat flux means the transport of energy by convection (moving air) and conduction into the atmosphere. The term sensible stems from this energy flux being associated with the temperature change of the air that we can sense. Latent heat flux is the energy flux related to a change of state, and refers here to the energy absorbed during the change of phase from liquid water (or ice) to water vapour. In other words, the latent heat flux in the above equation is the evaporation rate as an energy flux.

The ratio between the latent heat and sensible heat fluxes is referred to as the Bowen ratio B, and it is given by

$$B = \frac{H}{\lambda E} \tag{6.3}$$

Recall from (6.1) that evaporation, and hence the latent heat flux, is proportional to the vapour pressure difference, e_s-e_a, between the water surface and the overlying air. Similarly, the sensible heat flux is proportional to the temperature difference between the temperature at the water surface and in the overlying air. Transfer of both latent and sensible heat is largely based on turbulent eddies which are induced as the winds are affected by the frictional resistance of the surface. These turbulent eddies are chaotic motions of air where the movement of air also has vertical components (Figure 6.2). Transfer of latent and sensible heat with the turbulent eddies can be considered as a diffusion process described by Fick's law. Now if the diffusivities are assumed to be the same for both latent and sensible heat transfer, the Bowen ratio can be expressed as

$$B = \gamma \frac{T_s - T_a}{e_s - e_a} \tag{6.4}$$

where γ [kPa ºC^{-1}] is the psychrometric constant, T_s [ºC] is the surface temperature, and T_a [ºC] is the air temperature. Despite its name the psychrometric constant is strictly not constant but its value is dependent on air pressure and temperature. In the following exercise the psychrometric constant is assumed to be constant.

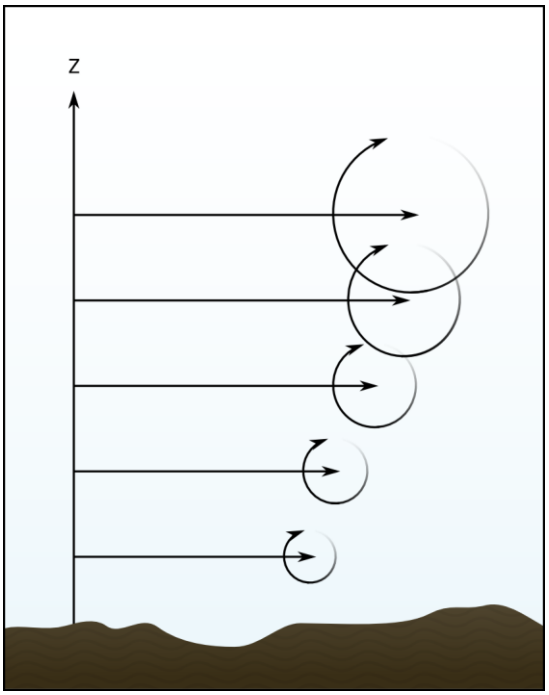

Figure 6.2 A schematic illustration of moving air in turbulent eddies. Due to friction at the ground the wind velocity increases with increasing distance from the ground (as depicted with longer arrows).

The slope of the saturation vapour pressure vs. temperature curve Δ [kPa ºC^{-1}] can be approximated as

$$\Delta = \frac{e_s^* - e_a^*}{T_s - T_a} \tag{6.5}$$

where e_s^* [kPa] is the saturation vapour pressure at the temperature of the evaporating surface and e_a^* [kPa] is the saturation vapour pressure at the air temperature. Solving for T_s – T_a from (6.5), inserting the result into (6.4), and noting that the vapour pressure at the evaporating surface e_s is equal to the saturation value e_s^* yields

$$B = \frac{\gamma\left(e_s^* - e_a^*\right)}{\Delta\left(e_s^* - e_a\right)} = \frac{\gamma\left(e_s^* - e_a^* - e_a + e_a\right)}{\Delta\left(e_s^* - e_a\right)} = \frac{\gamma}{\Delta}\left(1 - \frac{e_a^* - e_a}{e_s^* - e_a}\right) \tag{6.6}$$

With aid of the Bowen ratio, (6.2) can be written as

$$R_n - G = \lambda E\left(1 + B\right) = \left(1 + \frac{\gamma}{\Delta}\right)\lambda E - \frac{\gamma}{\Delta}\left(\frac{e_a^* - e_a}{e_s^* - e_a}\right)\lambda E \tag{6.7}$$

If the temperature of the evaporating surface is not known the saturated vapour pressure at the evaporating surface e_s^* is also not known. From Dalton's law (6.1) it follows

$$\frac{\lambda E}{e_s - e_a} = f \tag{6.8}$$

where f is the coefficient of proportionality. Inserting (6.8) into (6.7), recalling again that e_s is equal to the saturation value e_s^* and solving for λE produces

$$\lambda E = \frac{R_n - G + \frac{\gamma}{\Delta} f\left(e_a^* - e_a\right)}{1 + \frac{\gamma}{\Delta}} = \frac{\Delta\left(R_n - G\right) + \gamma f\left(e_a^* - e_a\right)}{\Delta + \gamma} \tag{6.9}$$

The above equation was first suggested by Penman (1948). Penman (1948) related the coefficient of proportionality to wind speed according to the following function

$$f_P\left(u_P\right) = 0.35\left(1 + 0.0098 u_P\right) \tag{6.10}$$

where f_P is the constant of proportionality in Penman's units, u_P is the wind speed at the height of two metres. The constant values in (6.10) require that the unit of evaporation is mm d^{-1}, the unit of wind speed is miles per day, and the unit of the vapour pressure deficit $e_a^* - e_a$ is millimetres of mercury. Inserting (6.10) into (6.9) and changing to more common units yields

$$E = \frac{\Delta\left(R_n - G\right)}{\lambda\left(\Delta + \gamma\right)} + \frac{\gamma}{\Delta + \gamma} \frac{6.43\left(1 + 0.536u\right)\left(e_a^* - e_a\right)}{\lambda} \tag{6.11}$$

where E [kg m^{-2} d^{-1}] is the rate of evaporation, Δ [kPa $ºC^{-1}$] is the slope of the saturation vapour pressure vs. temperature curve [-], R_n [MJ m^{-2} d^{-1}] is the net radiation, G [MJ m^{-2} d^{-1}] is the heat conduction into the lake water body, γ [kPa $ºC^{-1}$] is the psychrometric constant, e_a^* is the saturation vapour pressure [kPa] at the air temperature measured at the height of two metres, e_a [kPa] is the ambient vapour pressure of air at the height of two metres, λ [MJ kg^{-1}] is the latent heat of vaporization of water, and u [m s^{-1}] is the wind speed at the height of two metres. Evaporation rate in units of kg m^{-2} d^{-1} can be converted to mm d^{-1} by dividing the rate of evaporation as mass flux (E) by the density of water (1000 kg m^{-3}), i.e. 1 kg m^{-2} d^{-1} = 1 mm d^{-1}. Note also that latent heat of vaporization of water, λ is slightly dependent on the surface temperature of water. However, in the following exercise it is treated as a constant.

Equation (6.11) is recommended for estimating open water evaporation e.g. in Shuttleworth (1992). Note that as the second term in (6.11) involves an empirical wind function, the units must be specified as given in the brackets above. The unit of the second term is naturally mm d^{-1} although it cannot be inferred by dimensional analysis. The reason for inconsistent units is the empirical relationship between the vapour pressure deficit, wind speed and evaporation rate that does not comply with dimensional principles.

The Penman equation is often referred to as the combination equation as it combines mass transfer and energy balance approaches to estimating evaporation. The main advantage in the combination equation is that temperature measurements are required only at one elevation, and the temperature at the evaporating surface is not necessary for the application of (6.11) – assuming that the term R_n – G can be estimated otherwise (see below).

Estimating water vapour pressure deficit

The saturated water vapour pressure at the air temperature can be approximated by

$$e_a^* = 0.611 \exp\left(\frac{17.3 T_a}{T_a + 237.3}\right) \tag{6.12}$$

where the unit of vapour pressure is kPa and air temperature is given in °C. The ambient water vapour pressure is the saturated vapour pressure multiplied by the relative humidity, and the water vapour pressure deficit is the difference between the saturated and ambient values.

The slope of the saturation vapour pressure vs. temperature curve Δ [kPa $^{o}C^{-1}$] can be estimated at a given air temperature from

$$\Delta = \frac{4098 e_a^*}{\left(237.3 + T_a\right)^2} \tag{6.13}$$

where again the unit of vapour pressure is kPa and air temperature is given in °C.

Estimating net radiation

Net short-wave radiation S_n can be computed from

$$S_n = S_t \cdot \left(1 - \alpha\right) \tag{6.14}$$

where S_t [MJ m^{-2} d^{-1}] is the short-wave radiation reaching the ground and α [-] is the albedo (see Figure 6.1). The intensity of upward long-wave radiation is a function of the surface temperature, and the intensity of downward long-wave radiation is a function of air temperature and cloud cover. Penman (1948) used the Brunt (1939) equation, which does not require surface temperature data, to approximate the intensity of net long-wave radiation. It can be written as

$$L_n = -\sigma T_a^4 \left(0.56 - 0.25\sqrt{e_a}\right)\left(0.1 + 0.9\frac{n}{N}\right) \tag{6.15}$$

where L_n [MJ m^{-2} d^{-1}] is the net long-wave radiation, σ is the Stefan-Boltzmann constant [4.903 x 10^{-9} MJ m^{-2} K^{-4} d^{-1}], T_a [K] is the air temperature in Kelvins, and n / N is the ratio between actual and possible hours of sunshine. The vapour pressure of the air e_a needs to be given in kPa.

Net radiation R_n is simply the sum of net short- and long-wave radiation components, i.e.

$$R_n = S_n + L_n \tag{6.16}$$

Estimating G

Data on the heat exchange within the lake water body is rarely available, and therefore it can easily be neglected. However, it is worth noting that in large lakes evolution of the heat storage significantly influences the timing of evaporation (Croley, 1992).

EXERCISE 6.1

Objective

The objective here is to estimate lake evaporation based on meteorological values, compare the results against measured values, and discuss the reasons leading to differences between measurements and estimated values.

Meteorological data

The meteorological data for this exercise were kindly provided by the Finnish Meteorological Institute and the GGI-3000 evaporation data by the Finnish Environment Institute. It is not permitted to use this data for purposes other than solving this exercise.

Background

You have a data set from lake Pääjärvi, which lies in southern Finland, has the area of 13.4 km^2 and the maximum depth of 85 m. The daily meteorological data are for the years 1971-1974. In the same period lake evaporation was measured daily with a GGI-3000 evaporation pan installed on the lake. In addition, there are monthly Class A pan evaporation measurements (pan installed on ground) available from Hämeenkoski located at the distance of one kilometre from the lake.

In spring the direction of heat exchange between the lake surface and the lake water body is towards the lake as energy is absorbed in heating the lake water, whereas in late summer and autumn energy is released as lake water slowly cools down. The heat exchange between the lake surface and the lake water body (G) is estimated based on measurements conducted at lake Pääjärvi in 1970 (Elomaa, 1977). Based on Elomaa's results G is given as a fraction of net radiation from May to August, and as an absolute value in September and October when net radiation levels are so low that estimating G as a relative value of net radiation seems impractical.

Task

Calculate lake evaporation for lake Pääjärvi for each day using the Penman formula (6.11) and the data given in the Excel workbook called CMWRA_6_1.xls (sheet Ex6.1).

On the right side of the sheet there are summary tables and graphs for each year and month based on your daily calculations and the measurement data.

There are several predefined names (see Section 2.1) that you can use if you wish. It is easy to get confused with units, so be careful to remain aware as to what units should be used. Finally answer the following questions:

1. Compare the results of your lake evaporation calculations with the measurements. Are the differences between the measured and computed values in your opinion tolerable on a daily basis? Are the differences tolerable on a monthly basis?
2. There are some days when the measured lake evaporation is negative. What might be the reason for this?
3. There is a large difference between measured (10.3 mm/d) and calculated lake evaporation on August 12, 1972. What do you think could cause this difference? Would you trust the measured or the computed value more? Why?
4. Examine when Class A (installed on ground) and GGI-3000 (installed on lake) measurements differ from each other the most. Based on the energy balance of an evaporating surface (see (6.2)) try to explain the reasons for the differences.
5. Data for estimating the heat exchange within the lake (G) are rarely available. How does your lake evaporation estimate change if you neglect it? Report your findings on a monthly basis.

References

Brutsaert, W. 2005. Hydrology – An Introduction. Cambridge University Press, New York.

Brunt, D. 1939. Physical and dynamical meteorology.

Croley, T.E., II. 1992. Long-term heat storage in the Great Lakes. Water Resources Research, 28, 69-81.

Elomaa, E. 1977. Pääjärvi representative basin in Finland: heat balance of a lake. Fennia 149. Helsinki: Geographical Society of Finland. ISSN 0015-0010.

Penman, H.L. 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London, Series A, 193, 120–145.

Shuttleworth, W.J. 1992. Evaporation, Chapter 4. In: Maidment, D.R. (ed) Handbook of Hydrology. McGraw-Hill, New York, 4.1.-4.53.

7 SNOW ACCUMULATION AND MELT

T. Kokkonen, H. Koivusalo

7.1 Energy index approach

THEORY

A snow model is an essential part of quantifying the hydrological cycle in cold regions. A significant proportion of the annual runoff in cold regions may occur during a period of just a few weeks in spring, arising from snowmelt. A typical purpose for snow modelling is to provide an estimate of snowmelt input to be used in streamflow forecasting. Streamflow forecasts are necessary for issuing flood warnings and making water regulation decisions. Snow modelling can also aid in assessing how the projected changes in climate affect the seasonal snow cover and streamflow.

The main objective is usually to produce daily snowmelt discharge series, but the model also produces an estimate of the water stored in the snow pack. The latter information can be compared against field measurements. This exercise discusses an application of a degree-day snow model, which requires only precipitation and air temperature as meteorological input data. Degree-day models are commonly used in operational applications as precipitation and air temperature data are widely available.

Figure 7.1 shows a schematic conceptualising the two main snow processes to be described, namely accumulation of snow and snowmelt.

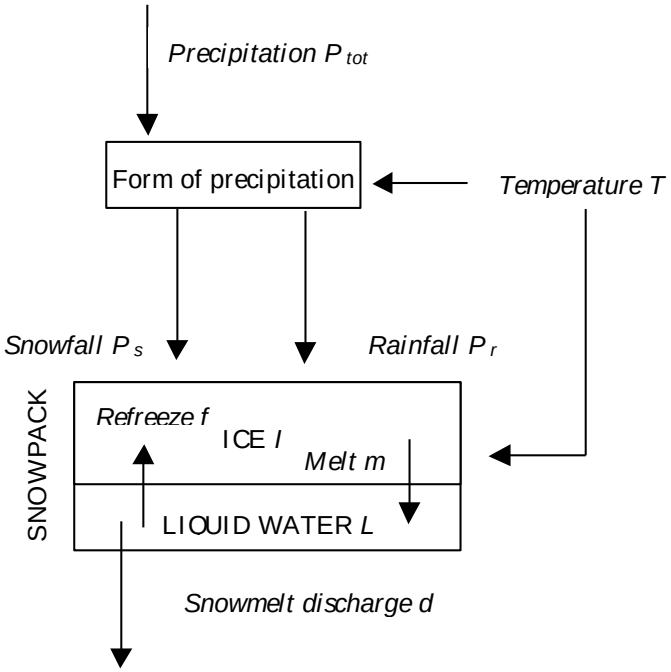

Figure 7.1. Schematic conceptualisation of snow processes.

A simple, conceptual snow model using degree days as the index for available energy requires as driving data

- daily precipitation [mm/d] over the area
- air temperature [°C] acting as a surrogate for the energy available for snowmelt

For this exercise, there are data available from an open site in Siuntio, southern Finland. Daily measurements of precipitation and air temperature cover the period extending from Dec 1, 1996 to Apr 30, 2000. For the same period there are snow water equivalent (SWE) measurements at intervals varying from 2 to 20 days (average 6 days). Precipitation was measured with a weighing gauge, and temperature was measured at two metres above the ground. Snow water equivalent was estimated from a matrix of 12 snow sticks, where snow depth was measured at each stick and snow density was measured at three locations.

The degree day snow model applied in this exercise is described mathematically as follows.

Form of precipitation

Below a certain threshold temperature $T_{\min}$ [ºC] all precipitation is assumed to fall in the form of snow, and above another threshold temperature $T_{\max}$ [ºC] all precipitation is in the form of rain. Between these two temperatures, the fraction of precipitation form is linearly interpolated and attains values between 0 and 1. In mathematical terms,

$$f_r = \begin{cases} 0 & T < T_{\min} \\ (T - T_{\min})/(T_{\max} - T_{\min}) & T_{\min} \le T \le T_{\max} \\ 1 & T > T_{\max} \end{cases}$$
$$f_s = 1 - f_r \tag{7.1}$$

where f_r is the fraction of rainfall, f_s is the fraction of snowfall, and T is the air temperature. Data from Iceland suggests that wet snowfall occurs between air temperature values of about 0.5 to 2.1 ºC (Ólafsson and Haraldsdóttir, 2003).

Correcting the precipitation for gauging error

For rainfall P_r [mm d^{-1}] and snowfall P_s [mm d^{-1}]

$$\begin{aligned} P_r &= c_r f_r P_{tot} \\ P_s &= c_s f_s P_{tot} \end{aligned} \tag{7.2}$$

where c_r [-] and c_s [-] are the correction coefficients for rainfall and snowfall, respectively, to correct for the proportions of rainfall and snowfall not registered by the gauge. P_{tot} is the gauged precipitation. According to Førland et al. (1996), the correction coefficients for measuring precipitation in an open site, depending on wind conditions at the gauge, range from 1.02 to 1.14 for rainfall, and from 1.05 to 1.80 for snowfall. In forested sites the effect of interception can be accounted for in the correction coefficients, which may lead to values below unity. Precipitation is typically gauged for open areas, so the snow that is intercepted in the tree canopy and subsequently evaporated has to be accounted for when modelling snow processes in forest areas.

Snowmelt

It is proposed that the rate of snowmelt is linearly related to the air temperature above the melting temperature. This can be written as

$$\begin{cases} m = k_d \left(T - T_{melt} \right) & T > T_{melt} \\ m = 0 & T \le T_{melt} \end{cases} \tag{7.3}$$

where m [mm d^{-1}] is the melt rate, k_d [mm °C^{-1} d^{-1}] is the degree-day factor for melt, and T_{melt} [°C] is the temperature where melting of snow is initiated. The value of T_{melt} is close to 0 °C, but can be allowed to differ slightly from it. Such a deviation can, for instance, account for a systematic difference between the air temperature at the measurement station and the temperature at the site where snow water equivalent is modelled. Air temperature has a strong relationship with altitude, and hence a difference in the elevation of the weather station and the modelling site easily results in a systematic bias of temperature measurements. Bergström (1990) reports that the degree-day factor for melt ranges from 1.5 to 4 mm °C^{-1} d^{-1} in operational streamflow forecasting applications in Sweden. According to Kuusisto (1984), the degree-day factor in a pine forest decreases from a value of 3.5 to 1 mm °C^{-1} d^{-1}, as the canopy density increases from 0 to 80%.

Freezing

Analogously to snowmelt, the rate of freezing f [mm d^{-1}] is written as

$$\begin{cases} f = k_f \left(T_{melt} - T \right) & T < T_{melt} \\ f = 0 & T \geq T_{melt} \end{cases} \tag{7.4}$$

where k_f [mm °C^{-1} d^{-1}] is the degree-day factor for freezing. The rate of freezing is lower compared to the rate of snowmelt and, therefore, the value of the parameter k_f is lower than k_d. Seibert (1997) and Seibert et al. (2000) reviewed the snow model structure of the HBV rainfall runoff model, where k_f was computed as a constant fraction of k_d. The fraction varied between 0.04 and 0.07.

Liquid water retention capacity of a snowpack

The liquid water retention capacity of a snowpack is related to the water equivalent of ice in the snow pack, i.e.

$$L_{\max} = rI \tag{7.5}$$

where L_{max} [mm] is the maximum amount of liquid water in the snow pack, r [-] is the retention parameter, and I [mm] is the water equivalent of ice in the snowpack. Tarboton and Luce (1996) referred to the retention parameter as the capillary retention fraction and fixed it to the value of 0.05.

Mass balance for the snowpack

The following equation gives the mass balance for the water equivalent of ice in the snowpack (see Figure 7.1).

$$\frac{dI}{dt} = P_s + f - m \tag{7.6}$$

The following equation gives the mass balance for the liquid water retained in the snow pack (see Figure 7.1).

$$\frac{dL}{dt} = P_r + m - f \qquad L \leq L_{\max} \tag{7.7}$$

When liquid water input (P_r + m - f) cannot fit into the liquid water store, i.e. the value of L exceeds L_{max}, the excess liquid water above L_{max} becomes snowmelt discharge d [mm d^{-1}].

Rain/melt

Rain/melt is snowmelt discharge when there is snow on the ground, and rainfall in snow-free periods. Rain/melt either infiltrates into the ground or ponds on the soil surface.

Parameter values

T_{max}, T_{melt}, c_s, k_d, k_f, and r are calibrated against measured SWE values. In order to reduce the number of calibration parameters T_{min} is set equal to T_{melt}. c_r is fixed to 1.05, according to the value suggested in Førland et al. (1996).

Model performance

The model is calibrated using the sum of squared errors (*SSE*) as an optimisation criterion:

$$SSE = \sum \left(y_i^{obs} - y_i^{sim} \right)^2 \tag{7.8}$$

where y_i^{obs} and y_i^{sim} are the observed and simulated, respectively, values at time step i. The calibration parameters are adjusted in an attempt to minimise the *SSE*.

The model performance is also evaluated in terms of the Nash – Sutcliffe (1970) efficiency

$$E_{NS} = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{obs} - y_i^{sim}\right)^2}{\sum_{i=1}^{N}\left(y_i^{obs} - \bar{y}^{obs}\right)^2} \tag{7.9}$$

where y_i^{obs} is the i^{th} observed SWE value, y_i^{sim} is the corresponding simulated value, $\bar{y}^{obs}$ is the mean observed value, and N is the number of observations. The maximum value of E_{NS} is 1, which indicates a perfect match between observed and modelled values.

Numerical solution

The ordinary differential equations (7.6) and (7.7) need to be solved numerically. Equation (7.6) can be discretized as follows:

$$\frac{I_{t+1} - I_t}{\Delta t} = P_{s,t+1} + f_{t+1} - m_{t+1} \tag{7.10}$$

where I_{t+1} is the water equivalent of ice at time t+1, Δt is the time step, $P_{s,t+1}$ is the snowfall at time t+1, f_{t+1} is the freezing at time t+1, and m_{t+1} is the melt at time t+1. Note that all fluxes at the right hand side of (7.10) are from time t+1. Solving I_{t+1} from (7.10) is known as the implicit Euler method.

Note that in the numerical solution you also need to take care that storages are not totally depleted, e.g. that melt does not exceed the amount of ice in the snow pack.

EXERCISE 7.1

Objective

The objective of this exercise is to learn how to solve numerically outflow from a linear store. Furthermore, the concept of an explicit versus implicit solution is also discussed.

Background

The outflow from a linear store is linearly related to its state – hence the term linear. This can be written as

$$Q_{out} = kX \tag{7.11}$$

where X is the state of the linear store, and k is the retention parameter. Now the mass balance for the linear store can be written as

$$\frac{dX}{dt} = Q_{in} - Q_{out} = Q_{in} - kX \tag{7.12}$$

where Q_{in} is the flow into the linear store and Q_{out} is the flow out of the linear store. Discretizing (7.12) yields the (explicit) solution

$$\frac{X^{t+1} - X^t}{\Delta t} = Q_{in}^t - kX^t \tag{7.13}$$

or the (implicit) solution

$$\frac{X^{t+1} - X^t}{\Delta t} = Q_{in}^{t+1} - kX^{t+1} \tag{7.14}$$

where X^t and Q_{in}^t are the state of the linear store and flow in to the store at time step t, respectively. As can be seen from (7.13) and (7.14), in an explicit solution the state variables (in this case X) are taken from the old time step t, whereas in an implicit solution the state variables are taken from the new time step $t+1$.

Task

Construct in Excel a linear store and solve outflow numerically using both explicit and implicit solution schemes. Compute the outflow from a linear store for the time period of 10 days. Use the following values.

- Q_{in} is always zero
- Initial value for X (X^0) is 100 mm
- k is 0.25 d^{-1}
- Δt is 0.5 d

Answer the following question:

1. In how long a time the state of the linear store drops to 50 mm?

EXERCISE 7.2

Objective

The idea is to investigate how predicted climate change affects seasonal snow cover. Table 7.1 shows predicted seasonal changes in temperature and precipitation. The change characterises the predicted difference in the climate variables between the future (2071 – 2100) and the control period (1961 – 1990) according to the IPCC (Intergovernmental Panel for Climate Change) B2 emission scenario (Nakicenovic and Swart, 2000) and a regional climate model application of the Swedish Meteorological and Hydrological Institute (Samuelsson, 2004).

Table 7.1. Predicted seasonal changes in temperature and precipitation between the future (2071 – 2100) and the control period (1961 – 1990).

	Temperature Change [ºC]	Precipitation Change [%]
Autumn (Sep-Nov)	2.9	2.0
Winter (Dec - Feb)	3.7	24.2
Spring (Mar - May)	3.3	-3.7
Summer (Jun - Aug)	1.3	-2.2

Task

Construct a degree day snow model and calibrate and validate it against measured SWE data from an open site in Siuntio, southern Finland. In workbook CMWRA_7_2.xls you will find a template for this exercise.

In the *Model* worksheet there are columns for different state variables (I and L) and fluxes (P_s, P_r, m, f). Study carefully the preceding theory and write in the worksheet cells equations that describe the dynamics of these fluxes/states. Apply the numerical scheme given in the theory section and use a time step of one day (Δt = 1 d) in the solution.

Note that the cells containing parameter values have been given names (*cr*, *cs*, *kd*, *kf*, *reten*, *tmax*, *tmelt*, *tmin*). You can use these names when typing Excel worksheet functions. The names correspond to the symbols used in the theory section – with the exception of the retention parameter (*reten* instead of just *r*).

Calculate in the last column (O) the daily computational mass balance error. Here the mass balance error is not the difference between observed and modelled values – it is merely an indicator of how well the model computationally maintains the mass balance, i.e. that the difference between incoming and outgoing fluxes equals the change in storage. This daily mass balance error should be very close to zero. If it is not, it is very likely that a mistake has occurred in constructing the numerical solution. Although a small computational mass balance error does not guarantee that a model is a good one, computing this error provides one good way to check if there are mistakes in construction of the numerical solution.

After you think that your model works technically correct, it is time to calibrate the parameter values. In model calibration, parameter values are adjusted in an attempt to achieve the best possible model fit against the measured SWE. Some initial guesses for the parameter values are provided in the *Model* worksheet, but the model fit can be much improved by calibrating the parameter values. See *Parameter values* in Section 7.1 for information about which parameter values you should calibrate.

Calibrate the parameter values using the *Solver* tool (*Tools -> Solver*) included in *Excel* (if you cannot find it you need to select *Tools -> Add-Ins* and to tick *Solver Add-In* from the list). SWE data from Dec 1996 to Apr 1999 are used for calibration, and the Nash – Sutcliffe efficiency (see equation (7.9)) is used as the objective function (cell O15 in *Model* worksheet). The workbook contains a custom function *NS* for computing the Nash – Sutcliffe efficiency. It is used as follows

=NS(*range1*, *range2*)

where *range1* and *range2* are cell ranges containing measured and modelled snow water equivalents, respectively.

In *Solver* the objective function is given as the *Target Cell*. You might want to apply some constraints on the values of the calibration parameters, e.g. that k_d and k_f may not be negative. This can be done in the *Solver* tool. Choose the initial values for parameters (i.e. the values that you type in the cells for parameter values before starting the *Solver* tool) according to the suggested ranges given in the theory section. In *SWE_Graph* worksheet you will find both SWE observations and the corresponding modelled values shown as a graph.

The model performance is assessed against SWE data not used in the calibration (from Nov 18, 1999, to Apr 14, 2000). Such assessment is typically referred to as validation of a model.

Based on your results, answer the following questions:

1. How do the calibrated parameter values correspond to values given in the literature?
2. How does validation performance compare with calibration performance?
3. What might cause the largest errors (e.g. 9 Dec 1999, 3 Mar 2000)?

Assume that measurement errors in precipitation and air temperature are 15% and 0.2 ºC, respectively. Which one of the errors has a larger impact on the annual maximum SWE?

Let us assume that climate changes according to the figures given in Table 7.1. Modify the available data from Dec 1996 to Apr 2000 to correspond with the projected changes, apply the model, and answer the following questions:

4. How does the length of the snow covered period change?
5. How does the maximum value of SWE change?
6. How does the time of occurrence of the maximum value of SWE change?

References

Bergström, S. 1990. Parametervärden för HBV-modellen i Sverige (in Swedish). SMHI Hydrologi. Nr 29, 1990.

Førland, E.J., Allerup, P., Dahlström, B., Elomaa, E., Jónsson, T., Madsen, H., Perälä, J., Rissanen, P., Vedin, H., and Vejen, F. 1996. Manual for operational correction of Nordic precipitation data. Report 24, Norwegian Meteorological Institute, 66 pp.

Kuusisto, E. 1984. Snow accumulation and snowmelt in Finland. Publications of the Water Research Institute 55, National Board of Waters, Helsinki, Finland, 149 pp.

Nakicenovic, N., and Swart, R. (Eds.). 2000. IPCC Special Report on Emissions Scenarios. Cambridge University Press, UK.

Nash, J.E., and Sutcliffe, J.V. 1970. River flow forecasting through conceptual models, Part I - A discussion of principles. Journal of Hydrology, 10, 282-290.

Ólafsson, H. and Haraldsdóttir, S.H. 2003. Estimation of the air temperature limits of snow and rain. Geophysical Research Abstracts, 5, 12798.

Samuelsson, P. 2004. RCAO and HadRM3p simulation results with focus on CLIME lake sites, WP2 Deliverable, SMHI Rossby Centre, SE-602 36 Norrköping, Sweden.

Seibert, J. 1997. Estimation of Parameter Uncertainty in the HBV Model. Nordic Hydrology, 28 (4/5), 247-262.

Seibert,, J., Uhlenbrook, S., Leibundgut, Ch., and Halldin, S. 2000. Multiscale calibration and validation of a conceptual rainfall-runoff model. Physics and Chemistry of the Earth, 25, 59-64.

Tarboton, D.G. and Luce, C.H. 1996. Utah Energy Balance Snow Accumulation and Melt Model (UEB), Computer model technical description and users guide. Utah Water Research Laboratory and USDA Forest Service Intermountain Research Station. http://www.neng.usu.edu/cee/faculty/dtarb/snow/snowreptext.pdf

8 RUNOFF

T. Kokkonen, H. Koivusalo

8.1 Rainfall-runoff modelling

THEORY

Historical background – origin of rivers

The origin of rivers was a subject of vigorous debate in the Aristotelian era. Aristotle discarded Plato's concept of a vast subterranean reservoir – Tartaurus – which was described as the source of all rivers and to which all rivers would flow back. Aristotle presented several arguments against the Tartaurus theory, and in particular he was concerned with replenishing the reservoir with new water to account for loss of water due to vaporisation. He did credit rainfall as one source of water flowing in rivers, but in his opinion rainfall alone would not be sufficient to feed all rivers of the world. As an additional supply of new water he proposed that there was a continuous conversion of air into water inside the earth, just like cold would change air to water above the earth.

> *Just as above the earth, small drops form and these join others, till finally water descends in a body of rain, so too we must suppose that in the earth the water at first trickles together little by little and that the sources of rivers drip, as it were, out of the earth and then unite.*
> Meteorologica*, Aristotle*

There is some basis to believe that Aristotle's pupil Theophrastus (371/370 – 288/287 B.C.) was the first person to have a correct perception of the hydrological cycle. Unfortunately, his original work on meteorology has been lost, but there is a four page long abstract that was composed in pursuance of translation of his work into Arabic. But it was not until the 17th century that it was shown in quantitative terms that precipitation alone would be sufficient to support the annual flow of rivers. Pioneering scientists advancing quantitative hydrology include the Frenchman Pierre Perrault (1608 – 1680), the Italian Edmé Mariotte (1620 – 1684), and the Briton Edmond Halley (1656 – 1742). Perrault showed in a catchment in the Seine basin that the annual rainfall was approximately six times greater than the annual flow in the river, thus indicating that precipitation indeed was capable of supporting the river flow. Mariotte did the same for a much bigger basin (the river Seine near Paris) and he also concluded that the annual rainfall was more than six times greater than the annual runoff. Halley's measurements showed that the amount of water evaporated from the oceans was clearly sufficient to sustain the flows of rivers.

Rainfall runoff modelling

Measurement-based evidence indicating that rainfall alone could sustain river flow laid a basis for searching for mathematical relationships between precipitation and streamflow. Probably the first serious attempts to estimate flood flow volumes can be traced back to a group of Irish engineers, who were given the task to design drainage channels in 1842 (Biswas, 1970). Ten years later, in 1851, Thomas Mulvaney (1822 – 1892) presented a paper which can be considered as the origin of the so-called rational method for flood peak estimation (Dooge, 1957). The method can be written as

$$q_{max} = f_r P_{max} A \tag{8.1}$$

where q_{max} [$m^3\ d^{-1}$] is the maximum rate of runoff, f_r [-] is the runoff coefficient, P_{max} [$m\ d^{-1}$] is the maximum rate of rainfall, and A [m^2] is the area of the catchment. The rational method has stood the test of time well as a simple tool to estimate peak flows.

It took 80 years before significant progress over the rational method was achieved in representing rainfall-runoff relationships mathematically. It came with the method proposed by Sherman (1932) which allowed calculation of 'continuous' hydrographs as opposed to merely delivering a peak flow estimate of the maximum flood event. Sherman called his method a unit-graph method, but it is presently better known as the unit hydrograph method. Mathematically the method can be written as

$$q(t) = \int u(\tau) r(t-\tau) d\tau \tag{8.2}$$

where t [d] is the time, q [m^3 d^{-1}] is the streamflow, u [m d^{-1}] is the effective rainfall (see below), and r [-] is the unit hydrograph. Effective rainfall is used instead of total rainfall as, mainly due to evapotransporative losses, not all of the rainfall is converted into streamflow. Effective rainfall is typically estimated using precipitation data together with other meteorological data, which are required to estimate the rate of evapotranspiration (e.g. see Croke and Jakeman (2004) for a range of effective rainfall models). Furthermore, Sherman thought the output of (8.2) to represent only the quick, surface runoff component. The so-called base flow, which is the part of a hydrograph present also between storm events, he attributed to discharge of groundwater.

As simulation of rainfall-runoff relationships has been a prime focus of hydrological research for several decades, an abundance of mathematical models has been proposed to quantify transformation of precipitation into streamflow. Following Beck (1991), mathematical rainfall-runoff models can crudely be classified as metric, conceptual and physics-based. Metric (black box) models are strongly observation-oriented, and they are constructed with little or no consideration of the features and processes associated with the hydrological system. The unit hydrograph method described above has a black box basis as the unit hydrograph simply attempts to reproduce the delay in the conversion of effective rainfall into streamflow, without really trying to answer how rainwater becomes streamflow.

Usually hydrologists have some kind of a perception in their mind about the behaviour of the hydrological system under study, and these *a priori* conceptions can be incorporated into a rainfall-runoff model. Conceptual models describe those component hydrological processes perceived to be of importance as simplified conceptualisations. This usually leads to a system of interconnected stores, which are recharged and depleted by appropriate component processes of the hydrological system. One of the first conceptual, hydrological models for continuous streamflow simulation was that of Linsley and Crawford (1960), which was developed to assess increase in the capacity of one of the water-supply reservoirs of the Stanford University. The structure of the model (Figure 8.1) is typical of many conceptual rainfall-runoff models, comprising storages representing upper level soil moisture, lower level soil moisture, groundwater, and water in the channel.

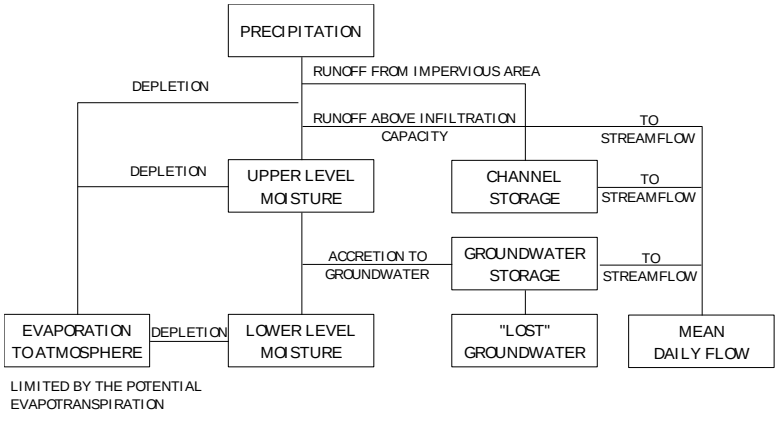

Figure 8.1. A flow diagram of the Linsley and Crawford model. Adopted from Linsley and Crawford (1960).

Finally, physics-based models attempt to mimic the hydrological behaviour of a catchment by using the concepts of classical continuum mechanics. Consequently, physics-based rainfall-runoff models are mostly spatially distributed – in contrast to lumped models which aggregate large entities (such as entire catchments) – and they split the modelling domain into several units (e.g. grid cells). Parameters of physics-based models should ideally be measurable in the field, but in practice some calibration of parameter values is nearly always necessary. One of the most often heard argument against physics-based hydrological models is related to the discrepancy between the measurement and computation scales. When physical model parameters applicable in some smaller scale (e.g. hydraulic conductivities measured from small cylindrical soil samples) are regionalised over the grid scale, they are commonly referred to as effective parameters. It can be argued that such effective parameters, which lump the effect of several small-scale processes into (a much larger) grid scale, are actually equivalent to parameters of lumped, conceptual models (see e.g. Beven, 1989; Beven, 1996). They cannot be directly measured in the field.

For further reading on rainfall-runoff models, see e.g. Kokkonen (2003).

Structure of a simple conceptual streamflow model

Figure 8.2 shows a schematic of the simple streamflow model to be constructed.

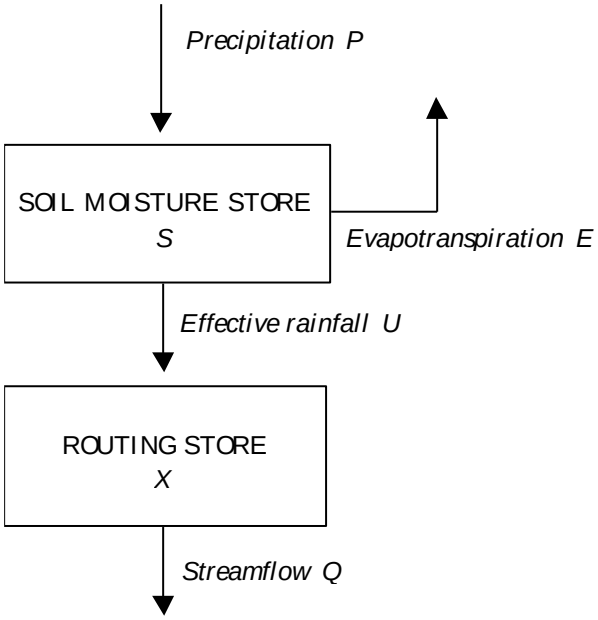

Figure 8.2. Schematic of the streamflow model.

The scheme for generating effective rainfall U is based on the structure of the HBV model reported in Bergström and Forsman (1973). Effective rainfall, meaning here the share of rain or melt that eventually emerges as streamflow, is computed from

$$U_t = \frac{S_t}{S_{MAX}} P_t \tag{8.3}$$

where U_t [mm d^{-1}] is the effective rainfall at time step t, S_t [mm] is the state of the soil moisture store at time step t, S_{MAX} [mm] is the capacity of the soil moisture store, and P_t [mm d^{-1}] is the rain/melt at time step t.

The soil moisture store is recharged by the rain/melt, and depleted by evaporation and effective rainfall. Its mass balance can be written as

$$\frac{dS}{dt} = P - E - U \tag{8.4}$$

where E [mm d^{-1}] is the evapotranspiration. Evapotranspiration is calculated from

$$E_t^P = \begin{cases} fT_t & ,T_t > 0 \\ 0 & ,T_t \le 0 \end{cases} \tag{8.5}$$

and

$$E_t = \frac{S_t}{S_{MAX}} E_t^P \tag{8.6}$$

where E_t^P [mm d^{-1}] is the potential evapotranspiration at time t, f [mm d^{-1} °C^{1}] is a calibration parameter, T_t [°C] is the air temperature, and E_t [mm d^{-1}] is the (actual) evapotranspiration. The above is a crude way of estimating potential evaporation but it is thought to be sufficiently accurate for the purposes of the current exercise. As parameter f is calibrated against streamflow data, the long-term sum of evaporation is assumed to become reasonably well estimated. The temperature dependence increases the potential evaporation in warm days when there is more energy available for evaporation. More detailed physics-based methods for estimating evaporation require data on radiation, wind speed, and relative humidity (see Section 6).

The delay between effective rainfall and streamflow on its route from the catchment to the flow gauging station is represented with a single linear routing store X. The outflow from X, i.e. streamflow, is defined as

$$Q = kX \tag{8.7}$$

where Q [mm d^{-1}] is the streamflow (in units of runoff) , k [d^{-1}] is the retention parameter, and X [mm] is the state of the routing store.

The mass balance for the routing store can be written as

$$\frac{dX}{dt} = U - Q \tag{8.8}$$

A model which is one step more complex than that of Bergström and Forsman (1973) is the IHACRES model. Jakeman and Hornberger (1993) present a slightly more complicated version of (8.6) for computing effective rainfall, and a routing store with two sub-stores in parallel (designated as a quick and a slow component of flow). They show that such a configuration is consistent with the amount of information in precipitation and streamflow data, inferring that more complexity increases the amount of non-uniqueness (see Section 4.5) of the estimated parameters.

Model performance

The model is calibrated using the sum of squared errors (*SSE*) as an optimisation criterion

$$SSE = \sum \left(y_i^{obs} - y_i^{sim} \right)^2 \tag{8.9}$$

where y_i^{obs} and y_i^{sim} are the observed and simulated, respectively, values at time step i. The calibration parameters are adjusted in an attempt to minimise the *SSE*.

The model performance is also evaluated in terms of the Nash – Sutcliffe (1970) efficiency

$$E_{NS} = 1 - \frac{\sum_{i=1}^{N} \left(y_i^{obs} - y_i^{sim} \right)^2}{\sum_{i=1}^{N} \left(y_i^{obs} - \overline{y}^{obs} \right)^2} \tag{8.10}$$

where y_i^{obs} is the i^{th} observed streamflow value, y_i^{sim} is the corresponding simulated value, $\bar{y}^{obs}$ is the mean observed value, and N is the number of observations. The maximum value of E_{NS} is 1, which indicates a perfect match between the observed and modelled values.

Numerical solution

The differential equations (8.4) and (8.8) are solved numerically in the same manner as in the snow exercise (Exercise 7.2). Equation (8.4) can be discretized as follows

$$\frac{S_{t+1} - S_t}{\Delta t} = P_{t+1} - E_{t+1} - U_{t+1} \tag{8.11}$$

Note that as both E and U are functions of S, you need to insert (8.3) and (8.6) into (8.11), and solve for S_{t+1}.

EXERCISE 8.1

Objective

Analogously to the snow exercise (Exercise 7.2) the idea is to investigate how projected climate change affects streamflow. Table 8.1 repeats projected seasonal changes in temperature and precipitation also used in the snow exercise in Section 7.1. The change characterises the projected difference in the climate variables between the future (2071 – 2100) and the control period (1961 – 1990) according to the IPCC (Intergovernmental Panel for Climate Change) B2 emission scenario (Nakicenovic and Swart, 2000) and a regional climate model application of the Swedish Meteorological and Hydrological Institute (Samuelsson, 2004).

Table 8.1. Predicted seasonal changes in temperature and precipitation between the future (2071 – 2100) and the control period (1961 – 1990).

	Temperature Change [ºC]	Precipitation Change [%]
Autumn (Sep-Nov)	2.9	2.0
Winter (Dec - Feb)	3.7	24.2
Spring (Mar - May)	3.3	-3.7
Summer (Jun - Aug)	1.3	-2.2

Task

Construct a simple conceptual, lumped streamflow model. The model is subsequently applied to the Palojärvenkoski basin (86.28 km^2, lake percentage 10%) located in Siuntio, in southern Finland (Figure 8.3). In workbook CMWRA_8_1.xls you will find a template for this exercise.

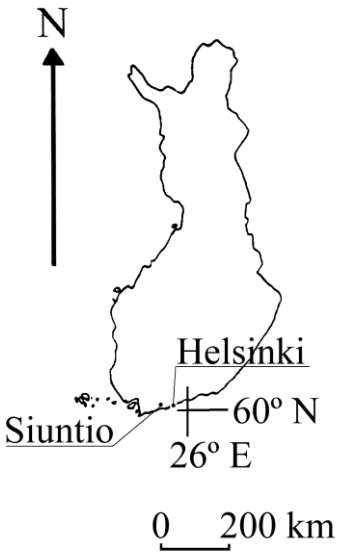

Figure 8.3. Location of Siuntio in Finland.

In the template worksheet you have columns for different state variables (S and X) and fluxes (E_p, U, Q). Note that the Rain/melt in the DATA section is the output of the snow model. Study carefully the preceding theory and write equations in the worksheet cells that describe the dynamics of these fluxes/states. Apply the numerical scheme given in the theory section and use a time step of one day (Δt = 1 d) in the solution (hint: using the unit time step means that it is not necessary to write the time step explicitly in the equations).

Note that the cells containing parameter values have been given names (*f*, *smax*, *k*). You can use these names when typing *Excel* worksheet functions. The names correspond to the symbols used in the theory section.

Calculate the daily computational mass balance error for the routing store X. Here the mass balance error is not the difference between observed and modelled values – it is merely an indicator of how well the model computationally maintains the mass balance, i.e. that the difference between incoming and outgoing fluxes equals the change in the storage. This daily mass balance error should be zero. If it is not, it is an indication of a mistake that has occurred in constructing the numerical solution. Although a small computational mass balance error does not guarantee that a model is a good one, computing this error provides one good way to check whether there are mistakes in construction of the numerical solution.

When you think that your model works technically correct, it is time to calibrate the parameter values. In model calibration the parameter values are adjusted in an attempt to achieve the best possible model fit against the measured streamflow.

Calibrate the parameter values using the Solver tool included in Excel (see Section 2.1). Flow data from Dec 1996 to Apr 1999 are used for calibration, and the Nash – Sutcliffe efficiency (equation (8.10)) is used as the objective function (cell I15). In Solver the objective function is given as the *Target Cell*. You might want to apply some constraints to the values of the calibration parameters, e.g. that k needs to be positive and S_{MAX} may not be negative. This can be done in the Solver tool. Starting values for parameter optimisation (parameters f, S_{MAX}, k), as well as initial values for the stores S and X, are given in the *Model* Excel sheet. In the *Streamflow_Graph* worksheet you will find both flow observations and the corresponding modelled values shown as a graph.

The model performance is assessed against flow data not used in the calibration (from May 1999 to Apr 2000). Such assessment is typically referred to as validation of a model, albeit always conditional as no model is ever true.

Based on your results, answer the following questions:

1. How does validation performance compare with calibration performance?
2. What might cause the error in Sept 1999 – Oct 1999?

Let us assume that climate changes according to the figures given in Table 8.1. Modify the model input data from Dec 1996 to Apr 2000 to correspond to the projected changes (you have the modified data available from the snow exercise, Exercise 7.2), apply the model, and answer the following questions:

3. How does the flow sum in the winter months from December to February change?
4. How do the flow peak values change?
5. How does the time of occurrence of the flow peak values change?
6. What is in your opinion the main mechanism behind the above changes?

8.2 Flood frequency analysis

THEORY

Flood frequency analysis uses probability distributions to assess how often flood events of a given magnitude occur, or alternatively to determine the magnitude of events having the same average recurrence interval or return period. Such analysis requires long periods of observed streamflow data. It is also necessary to assume that the hydrological behaviour of the river basin has not changed over the data record length. The technical term for this assumption is stationarity of the hydrological response, which means that the probability distribution of flows has not changed over the time period of observation. Then all streamflow values can be treated as random variables drawn from the same underlying probability distribution.

Return period

Of course streamflow events do not follow a regular pattern where flood peaks of similar magnitude occur at fixed time intervals. In flood frequency analysis the concept of a return period refers to the average time period between flood events of the same magnitude. If a flood of a given magnitude q is exceeded on average once in T years the probability of exceedence for any single year can be written as

$$P(Q > q) = \frac{1}{T} \tag{8.12}$$

where Q is the annual maximum flow. Note that for the sake of matching units on both sides of (8.12) T denotes the number of years without the unit of time. The cumulative probability value of the annual maximum flow $F(Q)$ gives the probability that a flood of a given magnitude is not exceeded

$$F(Q) = P(Q \le q) \tag{8.13}$$

As the probability of non-exceedence is the compliment for the probability of exceedence we can write

$$F(Q) = P(Q \le q) = 1 - P(Q > q) = 1 - \frac{1}{T} \tag{8.14}$$

Equation (8.14) expresses the relationship between the cumulative probability value and the return period of a given flood magnitude. Figure 8.4 is an example plot of a cumulative probability distribution function and demonstrates how the return period is determined when the cumulative probability value is known. In the example shown in Figure 8.4 the cumulative probability for the annual maximum flow value of 1250 m^3/s is 0.73. This means that in any year the annual maximum flow is less than or equal to 1250 m^3/s with probability 0.73. Or phrased another way, on the average in 73 out of 100 years the annual maximum flow does not exceed the value of 1250 m^3/s. Solving for the return period T from (8.14) yields

$$T = \frac{1}{1 - F(Q)} \tag{8.15}$$

And inserting 0.73 in (8.15) yields 3.7 years as the return period for the 1250 m^3/s flood event.

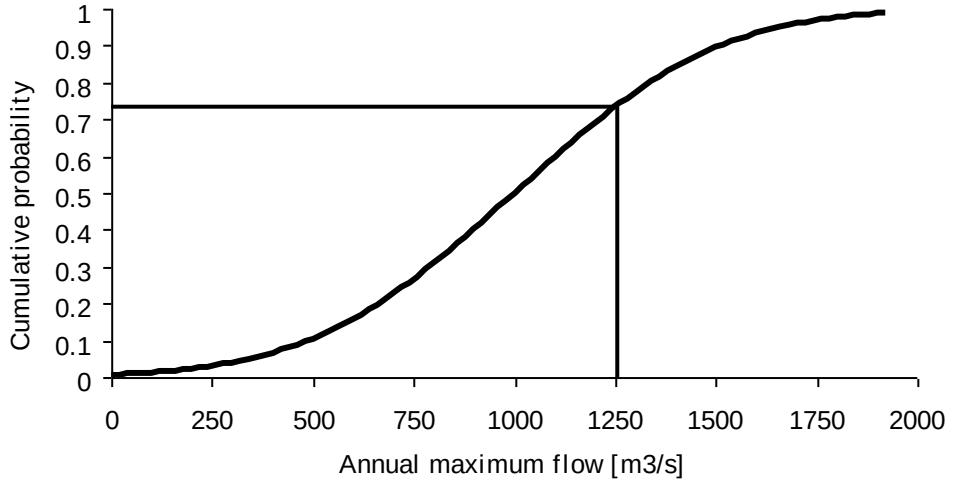

Figure 8.4. A cumulative probability distribution function of annual maximum flow.

Plotting position

Knowing the cumulative probability distribution of annual maximum flows allows for computation of the return period of a flood of the given magnitude (or vice versa, the magnitude of a flood with a given return period). While the true probability distribution of the maximum flow data is not known one can obtain an estimate of it based on the observed flow record. Preferably the flow record should cover a long period of time and the stationarity assumption holds so that the hydrological behaviour of the river basin should not have undergone any significant natural or manmade changes in that period. The former requirement assures that there is a sufficiently large sample of annual maximum flow values for credible estimation of the underlying probability distribution, and the latter requirement is essential for considering the maximum annual flow values of separate years as realisations from the same probability distribution.

Several formulae have been suggested for estimating the cumulative probability values of those annual maximum flows present in the available data. Estimation is typically based on ranking the flow events in the order of magnitude. In flood frequency analysis these estimated cumulative probability values are often referred to as plotting positions, owing to their use in probability plots. Perhaps the most straightforward way is that proposed by Hazen (Hazen, 1914). He suggested the probability scale be divided into as many equally spaced intervals as there are observations. The plotting positions are then computed from

$$F(Q_i) = \frac{i - 0.5}{N} \tag{8.16}$$

where Q_i is the i^{th} observation in the ascending order and N is the number of observations (Figure 8.5).Combining (8.15) and (8.16) gives $2N$ as the return period for the largest event in the record, which may seem unreasonable. Several plotting position formulae are listed e.g. in Rao and Hamed (2000) and in Stedinger et al. (1992).

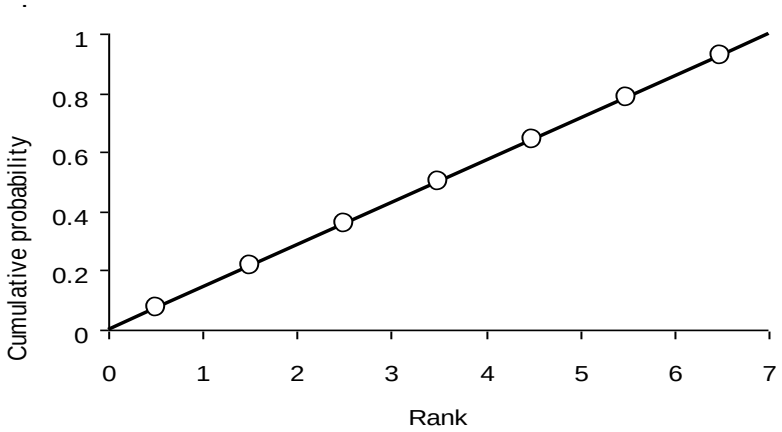

Figure 8.5. A schematic illustrating Hazen's plotting positions for a data record with seven[10] observations.

Gumbel probability distribution

Plotting positions provide only point estimates of the cumulative probability distribution for those flow values that are present in the observation data. Interpolation between the observed flow magnitudes and extrapolation outside the observed flood values are, however, often needed. In particular when determining the magnitude of rarely occurring design floods, which typically is the purpose of flood frequency analysis, extrapolation beyond the length of the observed record is necessary.

Fitting a theoretical probability distribution function to the available flow data provides the means for estimating cumulative probability values also for those flood magnitudes that are not included in the observed record. A large number of different probability distributions has been proposed for use in flood frequency analysis. National hydrological agencies have adopted from the numerous alternative distributions such distribution that is believed to deliver reasonably accurate estimates of hydrological risk for a nationwide use. Applying to each data set a different probability distribution according to the best possible fit would be too sensitive to sampling variations in the data (Stedinger et al., 1992). The Gumbel distribution has been popular in Finland and it will be used as an example in the following discussion. It is noted, however, that the discussion would be quite analogous if some other probability distribution had been studied.

Assume that the annual maximum flow value Q is a random variable following the Gumbel distribution. Its cumulative probability is then given by

$$F(Q) = \exp\left(-\exp\left(-\frac{Q-\beta}{\alpha} \right) \right) \tag{8.17}$$

where α and β are parameters determining, respectively, the scale and location of the distribution. Figure 8.6 illustrates the shape of the probability density function for a Gumbel distributed random variable.

[10] Seven observations are of course too few for flood frequency analysis. However, the idea of plotting position values as a function of the rank becomes clearer when the graph is drawn with just a small number of data.

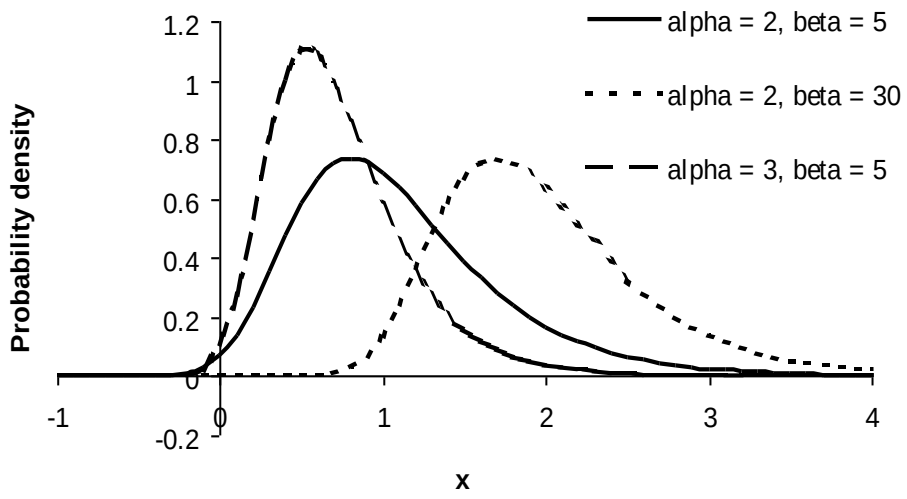

Figure 8.6. Probability density function of the Gumbel distributed random variable *x* for different α and β parameter values.

Parameter estimation

Before a theoretical probability distribution can be applied to estimating flood frequencies its parameters need to be estimated. Parameter estimation here means adjustment of parameters in such a way that the theoretical probability distribution yields the best possible description of the observed flood frequencies (see also Section 3.1). Two commonly used methods for parameter estimation in flood frequency studies are the method of moments and the method of maximum likelihood. In the method of moments the sample moments derived from the flow observations are equated with the corresponding population moments of the theoretical probability distribution. The number of moments to be considered is equal to the number of parameters in the theoretical probability distribution. For Gumbel distribution having two parameters the first two moments, i.e. the mean and the variance, are needed. The mean of the Gumbel distribution is given by

$$\beta + 0.5772157\alpha \tag{8.18}$$

and the variance by

$$\left(\frac{\pi^2}{6}\right)\alpha^2 \tag{8.19}$$

In the maximum likelihood method parameter values are identified such that the probability of the studied distribution generating the observed flood values attains its maximum. A brief description of the maximum likelihood estimation with illustrating examples can be found at http://en.wikipedia.org/wiki/Maximum_likelihood.

Probability plot

The adequacy of a theoretical distribution in generating the observed flow data can graphically be evaluated with the aid of a probability plot, where flow observations are plotted against the flow values estimated from the theoretical probability distribution. If the theoretical probability distribution exactly describes the probability of occurrence for the observed flow magnitudes, then the probability plot would appear as a straight line through the origin with a 45° slope. Solving for the annual maximum flow value *Q* from Gumbel's cumulative probability distribution (8.17) yields

$$Q = \beta - \alpha \ln\left(-\ln F\right) \tag{8.20}$$

Equation (8.20) is the key to constructing a probability plot for the Gumbel distribution. Plotting observed annual maximum flow values against $\beta - \alpha \ln(-\ln F)$ yields a one-to-one relationship when the occurrence of flood events exactly follows the Gumbel probability distribution and the cumulative probability value for each annual maximum flow observation is known. Of course in reality such a perfect match is never achieved. Flood observations do not perfectly follow any theoretical probability distribution and cumulative probability values derived from plotting position formulae are only estimates. The deviations of the probability plot points from the straight line, however, help in judging the suitability of the Gumbel probability distribution to the observed flood frequencies (Figure 8.7).

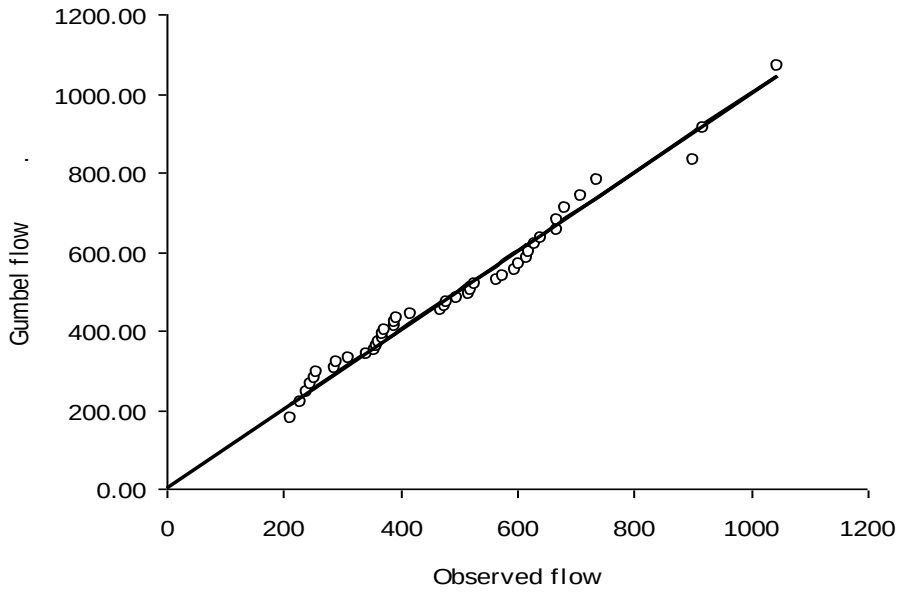

Figure 8.7. An example of a probability plot.

Confidence intervals

The uncertainty arising from estimating parameter values of the selected theoretical probability distribution from a limited number of observations can be assessed by constructing confidence intervals for the computed flood magnitude estimates (see *Confidence intervals* in Section 4.1). When the Gumbel distribution parameters have been estimated using the method of moments the standard deviation of the flood magnitude estimate S_G can be expressed as

$$S_G = \sqrt{\frac{\alpha^2}{N}\left(1.15894 + 0.19187Y + 1.1Y^2\right)} \tag{8.21}$$

where Y is the reduced Gumbel variate defined by

$$Y = -\ln\left(-\ln F\right) \tag{8.22}$$

It can be shown that when the number of observations increases Gumbel's distribution approaches the normal distribution. Therefore an approximate $(1 - \alpha)$ confidence interval can be computed from

$$Q_G \pm z_{\alpha/2} S_G \tag{8.23}$$

where z is the normal standard variate.

It is worth stressing that the confidence interval described above only accounts for the uncertainty stemming from estimating the probability distribution parameters from a data record of limited size[11]. The incapability of the selected theoretical probability distribution (here the Gumbel distribution) to describe the flood reoccurrence, however, may at longer return periods in particular result in much larger errors than the uncertainty in parameter estimates. While flood frequency analysis has been widely used worldwide in delivering design flood values, the validity of the assumptions on which it is based has also received much criticism and questioning (e.g. Klemeš, 1986).

Further reading

Readers interested in flood frequency analysis are referred to a comprehensive text book by Rao and Hamed (2000).

EXERCISE 8.2

Objective

The objective of this exercise is to learn how flood frequency analysis can be used to deliver design flood estimates of a given recurrence interval. The Gumbel probability distribution is used as an example.

Streamflow data

The streamflow data for this exercise were kindly provided by the Finnish Environment Institute. It is not permitted to use this data for purposes other than solving this exercise.

Task

You have 43 years of streamflow data from River Ivalonjoki located in the Finnish Lapland. Your task is to deliver an estimate of a 100 year flood for this river. In workbook CMWRA_8_2.xls you will find a template for this exercise. Proceed as follows:

- Fit the Gumbel distribution to the observed annual peak flows.
- Construct a probability plot for visually inspecting how well the observed maximum flows follow the Gumbel distribution. Use the Gringorton plotting position given by

$$F(Q_i) = \frac{i - 0.44}{N + 0.12}$$

  This is the recommended plotting position formula for the Gumbel distribution.
- Construct 95% confidence intervals in the probability plot.
- Produce the estimate for the 100 year flood together with its 95% confidence interval.

Finally answer the following questions:

1. In your mind, how well does the Gumbel distribution fit the observed annual maximum flows?
2. How many of the observed maximum flows fall outside the 95% confidence intervals of the Gumbel distribution?
3. What is the probability that the 100 year flood occurs the next year?
4. What is the probability that the 100 year flood does not occur at all in the next 100 years?

[11] The parameter values are thus sample estimates. See also footnotes 3 and 4.

Hints

The *INDIRECT* Excel worksheet function (see Section 2.1) is useful for finding the annual maximum values from daily records of streamflow. It is relatively easy to define the first and last rows for each year, but you must take leap years into account, and you need to check for missing days in the data. This is easily accomplished by verifying that the first date is always Jan 1, and the last date is Dec 31. Now all you require is a formula in the *Max. Discharge* column for finding the maximum value of a cell range C[first row]:C[last row], which you can write using INDIRECT and MAX functions.

References

Beck, M.B. 1991. Forecasting environmental change. Journal of Forecasting 10, 3-19.

Bergström, S., and Forsman, A. 1973. Development of a conceptual deterministic rainfall-runoff model. Nordic Hydrology, 4, 147-170.

Beven, K. 1989. Changing ideas in hydrology: the case of physically based models. Journal of Hydrology, 105, 157-172.

Beven, K.J. 1996. A discussion of distributed hydrological modelling. In: Abbott, M.B., Refsgaard, J.C. (Eds.), Distributed Hydrological Modelling. Water Science and Technology Library. Kluwer Academic Publishers, Dordrecht, pp. 255-278.

Biswas, A.K. 1970. History of hydrology. North-Holland Publishing Company, Amsterdam, 336 pp.

Croke, B.F.W., and Jakeman, A.J. 2004. A catchment moisture deficit module for the IHACRES rainfall-runoff model. Environmental Modelling and Software, 19, 1-5.

Dooge, J.C.I. 1957. The rational method for estimating flood peaks. Engineering 184, 311-313; 374-377.

Hazen, A. 1914. Discussion on 'Flood flows' by W.E. Fuller, Transactions of the American Society of Civil Engineers, 77, 526-632.

Jakeman, A.J., and Hornberger, G.M. 1993. How much complexity is warranted in a rainfall-runoff model? Water Resources Research, 29, 2637-2649.

Klemeš, V. 1986. Dilettantism in hydrology: Transition or destiny? Water Resources Research, 22, 177S–188S.

Kokkonen, T. 2003. Rainfall-runoff modelling – comparison of modelling strategies with a focus on ungauged predictions and model integration. Helsinki University of Technology Water Resources Publications, TKK-VTR-9, Espoo, Finland.

Linsley, R.K., Crawford, N.H. 1960. Computation of a synthetic streamflow record on a digital computer. Publication 51, International Association of Scientific Hydrology, Stanford, 526-538 pp.

Nakicenovic, N., and Swart, R., (Eds.). 2000. IPCC Special Report on Emissions Scenarios, United Kingdom, Cambridge University Press.

Nash, J.E., and Sutcliffe, J.V. 1970. River flow forecasting through conceptual models, Part I - A discussion of principles. Journal of Hydrology, 10, 282-290.

Rao, A.R., and Hamed, K.H. 200. Flood Frequency Analysis, Boca Raton, the United States, CRC Press, 350 pp.

Samuelsson, P. 2004. RCAO and HadRM3p simulation results with focus on CLIME lake sites, WP2 Deliverable, SMHI Rossby Centre, SE-602 36, Norrköping, Sweden.

Sherman, L.K., 1932. Streamflow from rainfall by the unit-hydrograph method. Engineering News Record 108, 501-505.

Stedinger, J.R. 1992. Frequency analysis of extreme events, Chapter 18. In: Maidment, D.R. (ed) Handbook of Hydrology. McGraw-Hill, New York, 18.1.-18.66.

9 GROUNDWATER

T. Kokkonen, H. Koivusalo, L. Stenberg

9.1 Steady state groundwater flow

THEORY

Definition of groundwater

Water beneath the ground is referred to as groundwater when it exists below the groundwater table where the pores in the soil are fully filled with water. The pressure at the groundwater table is atmospheric. This is also how the groundwater table is defined; the groundwater table is the level at which the pressure in water is atmospheric. The layer just above the groundwater table may be filled (saturated) with water due to capillary forces, and consequently it is called the capillary fringe. Perhaps the easiest way to understand what a groundwater table means is to think of it as the level where water settles in a well or in a drill hole. Often the atmospheric pressure is used as a reference and is given the value of zero. In other words, any pressure value being greater than atmospheric is positive, and correspondingly any pressure being smaller than atmospheric is negative.

Hydraulic head

Hydraulic head is defined as the amount of energy per unit weight of water. As velocities in groundwater flow are typically slow, kinetic energy is disregarded and energy can be considered to be composed of gravitational and pressure energy. Potential energy due to gravitation E_g [J] at a given point can be expressed as

$$E_g = mgz \tag{9.1}$$

where m [kg] is the mass of water, g is the acceleration due to gravity [m s^{-2}], and z [m] is the elevation of the point. The energy of water due to its pressure E_p is defined as

$$E_p = pV \tag{9.2}$$

where p [Pa] is the pressure of water and V [m^3] is the volume of water. The amount of energy required for moving from the reference point to a given point can be computed from the Bernoulli equation (note: mass = volume x density => volume = mass / density)

$$E = \int_{z_o}^{z} mgdz + \int_{p_0}^{p} \frac{m}{\rho} dp \tag{9.3}$$

where z_0 and p_0 are the elevation and pressure at the reference point, respectively, and ρ [kg m^{-3}] is the density of water. Now if it is assumed that water is incompressible, and that the values for elevation and pressure at the reference point are zero, (9.3) takes the form

$$E = mgz + \frac{m}{\rho} p \tag{9.4}$$

Dividing the energy E by the weight of water mg yields the hydraulic head H [m] at the given point

$$H = z + \frac{p}{\rho g} \tag{9.5}$$

The head component due to gravity z [m] is called gravity head, and the head component due to pressure $\frac{p}{\rho g}$ is correspondingly referred to as pressure head. Pressure head is typically denoted by the symbol h [m], and then (9.5) becomes

$$H = z + h \tag{9.6}$$

Figure 9.1 illustrates the distribution of hydraulic head into gravity and pressure head in a hydrostatic container (i.e. in a bucket of standing water).

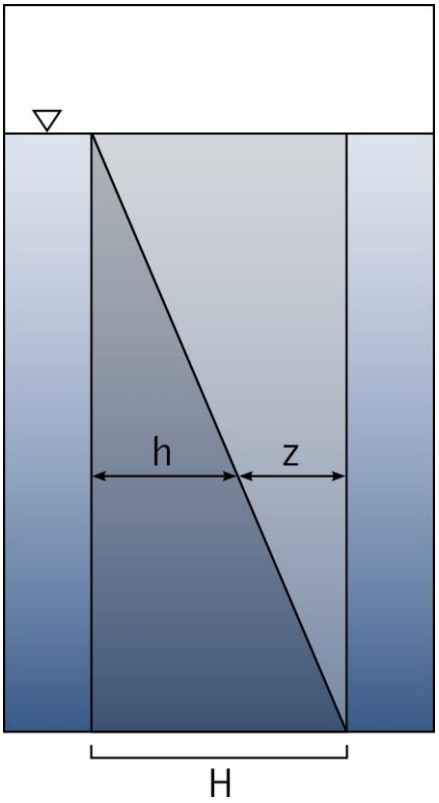

Figure 9.1. Distribution of hydraulic head (H) into gravity head (z) and pressure head (h) in a hydrostatic container. Elevation is measured from the bottom of the container (i.e. elevation is zero at the bottom).

Darcy's law

The work of the French engineer Henry Darcy established that the rate of water flow in a sand filled column was directly proportional to the difference in the hydraulic head and to the cross-sectional area of the column, and inversely proportional to the length of the sand filter. This finding is known as Darcy's law, and it can be written as

$$Q = \frac{KA\Delta H}{L} \tag{9.7}$$

where Q [$m^3\ s^{-1}$] is the discharge, K [$m\ s^{-1}$] is the proportionality constant (*hydraulic conductivity*), A [m^2] is the cross-sectional area, ΔH [m] is the change in the hydraulic head, and L [m] is the length of the sand column. Dividing (9.7) by the cross-sectional area, considering that the direction of flow is towards a smaller head, and letting the length L (in the direction of x) become infinitesimally small, gives

$$q = -K\frac{dH}{dx} \tag{9.8}$$

where q [m s^{-1}] is discharge per unit cross-sectional area (specific discharge) and x [m] is a spatial coordinate.

Governing equation for groundwater flow

Darcy's law, which is in essence a momentum balance equation, needs to be complemented with the mass balance equation to obtain the governing equation for groundwater flow. Let us consider a steady-state situation, which means that the flow rate does not change in time in any given location within the aquifer. Steady state also implies that the amount of water present in any given location in the aquifer remains constant through time.

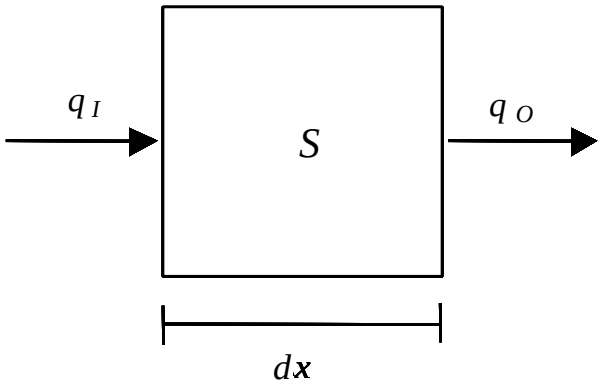

Figure 9.2. One-dimensional flow through a control box.

Figure 9.2 depicts one-dimensional flow through a control box. In steady-state the state of the storage S does not change with time, and the outflux q_O must equal the influx q_I. Hence we can write

$$\frac{dq}{dx} = 0 \Rightarrow -\frac{dq}{dx} = 0 \tag{9.9}$$

where q is the flux and x is the location coordinate. The term $-\frac{dq}{dx}$ can be verbally expressed as the excess of influx over outflux, and in the steady-state case, where no water is stored within the control box, it needs to equal zero. Now combining Darcy's law and the mass balance equation, i.e. (9.8) and (9.9), yields

$$-\frac{dq}{dx} = -\frac{d}{dx}\left(-K\frac{dH}{dx}\right) = \frac{d}{dx}\left(K\frac{dH}{dx}\right) = 0 \tag{9.10}$$

Equation (9.10) is the governing equation for one-dimensional, steady-state groundwater flow. In a homogeneous case, i.e. when the hydraulic conductivity K does not change in space, (9.10) reduces to

$$\frac{d^2H}{dx^2} = 0 \tag{9.11}$$

which is also known as the Laplace equation.

Sometimes it is necessary to account for situations where water is injected into – or extracted from – the aquifer. This can be handled with the so-called sink/source term, which is the amount of the extracted/injected water per unit area and unit time. In Figure 9.3 the double arrow depicts a sink or source within an aquifer. The mass balance must also hold in this case, which allows us to write

$$\frac{dq}{dx}\Delta x\left(b\Delta y\right) = R\Delta x\Delta y \tag{9.12}$$

where Δx [m], Δy [m], and b [m] are dimensions of the control box, q [m s^{-1}] is the flux, and R [m s^{-1}] is the sink/source term. In this case the extracted/injected amount of water must match the change in flux. Let us take a closer look at various terms of (9.12)

- $\frac{dq}{dx}\Delta x$ is the change in flux over the distance Δx
- $b\Delta y$ is the area of the face of the box in the direction of flow
- $\frac{dq}{dx}\Delta x\left(b\Delta y\right)$ is the difference between the rate of inflow and outflow
- $R\Delta x\Delta y$ is the volume of the extracted/injected water per unit time

Clearly (9.12) is stating that the volume of the extracted/injected water is reflected as a difference between inflow to the control box and outflow from the control box.

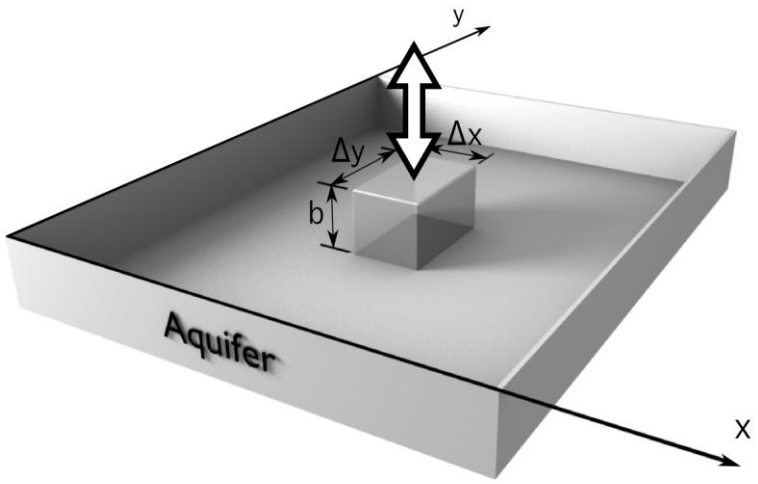

Figure 9.3. Schematic for the mass balance in presence of sinks or sources.

Dividing by $\Delta x\Delta y$ and inserting Darcy's law (equation (9.8)) for q let us rewrite (9.12) as

$$\frac{d}{dx}\left(-Kb\frac{dH}{dx}\right)=R \tag{9.13}$$

In a homogeneous case (9.13) reduces to

$$\frac{d^2H}{dx^2}=-\frac{R}{Kb}=-\frac{R}{T} \tag{9.14}$$

where T [m^2 s^{-1}] is the transmissivity of the aquifer. Transmissivity is defined to be the product of the hydraulic conductivity K and the aquifer thickness b, and it describes the capability of the aquifer to convey water.

Boundary conditions

A governing equation alone is not enough to describe a specific groundwater problem, but some supplementary information is required. Boundary conditions describe how the area of interest interacts with the surrounding areas. There are two main types of boundary conditions

1) Constant (or prescribed) head boundary – Dirichlet boundary condition
2) Constant (or prescribed) flux boundary – Neumann boundary condition

As the names suggest, in the first case the hydraulic head at the boundary is known, and in the second case the flux through the boundary is known. A common special case to the constant flux boundary condition is an impervious boundary (i.e. the flux through the boundary is zero).

Numerical solution

Analytical solutions to the governing equation of groundwater flow exist in some special cases, but usually it needs to be solved numerically. In order to solve (9.14) numerically using the method of finite differences, the equation first needs to be discretized. In a discretized form (9.14) can be written as

$$\frac{\left[\frac{H_{i+1}-H_i}{\Delta x}-\frac{H_i-H_{i-1}}{\Delta x}\right]}{\Delta x}=-\frac{R_i}{T} \tag{9.15}$$

where H_i is the hydraulic head at node i, and R_i is the source/sink term at node i (Figure 9.4).

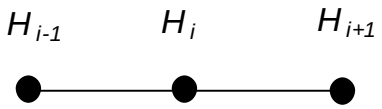

Figure 9.4. Hydraulic gradient at node i (H_i), and at preceding and following nodes (H_{i-1} and H_{i+1}).

Numerical representation of the constant head boundary condition is straightforward: a calculation node having a constant head boundary is simply assigned with the value of the constant hydraulic head.

An impervious boundary can be represented with an imaginary node that mirrors the head value of the node on the other side of the boundary. Hence the hydraulic gradient is forced to equal zero and no flow occurs through the boundary.

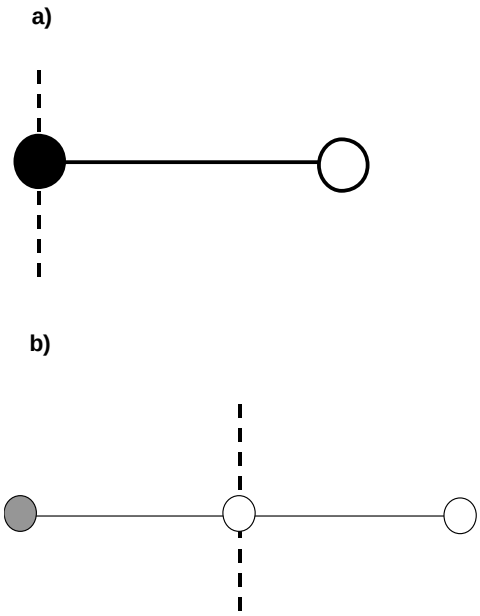

Figure 9.5. Numerical representation of a constant head boundary (a) and an impervious boundary (b). The dashed line shows the location of the boundary. In (a) the black node is assigned the known constant head value. In (b) the imaginary (grey) cell is forced to have the same value as the node on the other side of the boundary.

Aquifer types

An aquifer is defined to be a geological formation which permits extraction of water from it in significant quantities. This means that an aquifer needs to contain water and the water has to be able to move through the aquifer at a sufficient velocity. The latter condition simply implies that the hydraulic conductivity of an aquifer may not be too low. Aquifers are typically classified as confined or unconfined aquifers, depending on how they are located in relation to confining, impervious – or close to impervious – layers. A confined aquifer, or a pressure aquifer, resides between two confining layers and therefore the water contained in the aquifer is under pressure. When a well is bored into the aquifer, the water level in the well will rise above the upper confining layer. In an unconfined aquifer the upper boundary is the free groundwater table. See Figure 9.6 for an illustration of a confined and an unconfined aquifer. Modelling movement of water in a confined aquifer is easier than in an unconfined aquifer, as in a confined aquifer the thickness of the saturated, water transmitting layer does not change with a changing water level.

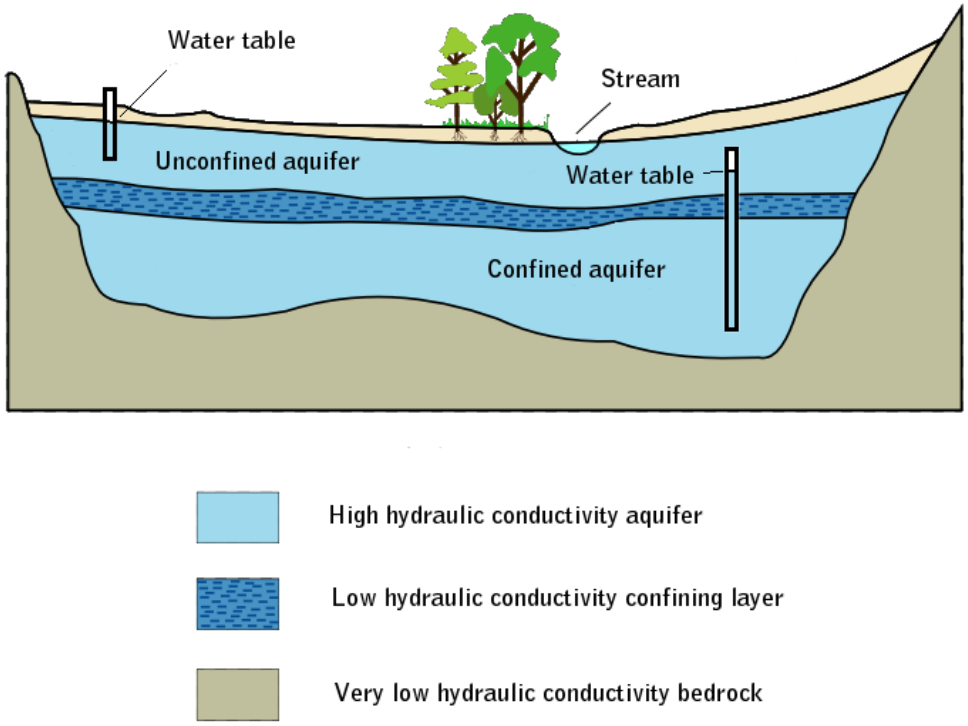

Figure 9.6. Schematic illustration of a confined and an unconfined aquifer. Groundwater tables in wells installed to the aquifers are also shown. Modified from http://upload.wikimedia.org/wikipedia/commons/0/04/Aquifer_en.svg.

Heterogeneity and isotropy

An aquifer is said to be homogeneous if its properties (e.g. hydraulic conductivity) are not dependent on the location. In the opposite case the aquifer is heterogeneous. And when the aquifer properties do not change with the direction of flow, the aquifer is isotropic – otherwise it is anisotropic. Figure 9.7 is a schematic illustrating both homogeneity and isotropy.

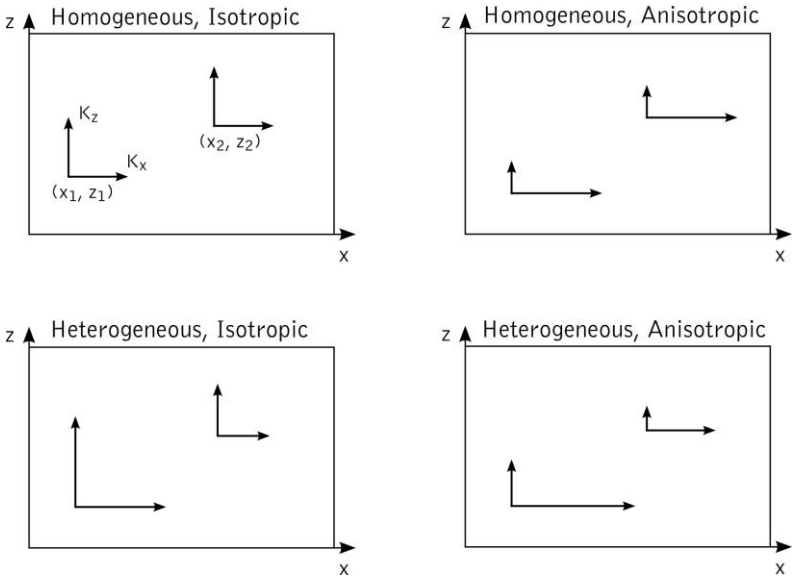

Figure 9.7. Schematic illustration of heterogeneity and isotropy. The longer the arrow the greater the hydraulic conductivity is in the direction of the arrow.

EXERCISE 9.1

Objective

This exercise has two objectives. The first objective is to understand how measured hydraulic head values can be used in determining the transmissivity of an aquifer. The second objective is to learn how a simple, numerical groundwater model can be constructed and solved.

Background

Hydraulic conductivities measured from soil samples having a relatively small size tend not to give a realistic estimate of the field-scale water conveyance capacity of an aquifer. Transmissivity of an aquifer, i.e. the product of an average hydraulic conductivity and aquifer thickness, can be inversely inferred from hydraulic head data obtained from a pumping test. The inverse inference means here that hydraulic heads at selected points are measured and the governing equation is solved for transmissivity.

In the early 20th century Thiem (1906) developed an analytical method for determining the average transmissivity from a steady-state pumping test. Although a steady-state pumping test may not be very practical, due to a potentially long pumping time and a large amount of water to be pumped, it serves well as a demonstration of how to determine an average transmissivity from hydraulic head data.

Figure 9.8 shows a schematic of the pumping test. A pumping well extracts water from the aquifer, and this causes the level of the hydraulic head to drop around the well. Darcy's law allows us to write

$$q_r = T\frac{dH}{dr} \tag{9.16}$$

where q_r [$m^3\ s^{-1}\ m^{-1}$] is the discharge per unit width, T [$m^2\ s^{-1}$] is the transmissivity, H [m] is the hydraulic head, and r [m] is the distance from the pumping well.

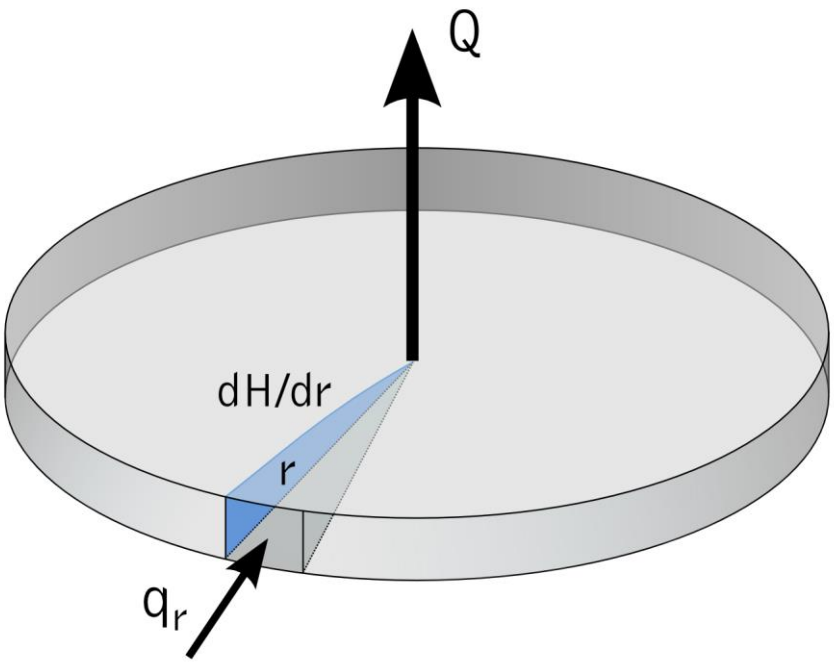

Figure 9.8. A schematic of a steady-state pumping test.

The discharge of the well, *Q*, must equal the discharge per unit width multiplied by the perimeter of the circle at distance *r*, i.e. in mathematical terms

$$Q = 2\pi r T \frac{dH}{dr} \tag{9.17}$$

The assumptions underlying Thiem's analytical solution are

1) The areal extent of the aquifer is infinite.
2) The aquifer is confined, homogeneous, isotropic, and has a constant thickness
3) Before the pumping commences, the surface of the hydraulic head is horizontal
4) Pumping rate does not change over time
5) The pumping well fully penetrates the aquifer.

In workbook CMWRA_9_1.xls you will find templates for the following exercises. **Note! You need to switch the iteration mode on in Excel, otherwise you will get error messages complaining about circular references!** For instructions on how to do this, see Section 2.1. Set *Maximum iterations* to 1000, and *Maximum change* to 0.00001.

Task

- **Exercise 9.1a**

In sheet Ex9.1a you have drawdown data for the pumping test. The drawdowns have been recorded after steady-state has been achieved. Drawdown is defined as the drop in hydraulic head from the initial level. In sheet Ex9.1a you also will find pairs of drawdown (s) and distance from the pumping (well). The discharge of the pumping well *Q* is also given.

Solve via the differential equation (9.17) for transmissivity *T* (you need to do this by integration). This equation is known as Thiem's equation.

1. Report derivation of Thiem's equation.
2. Use Thiem's equation to estimate transmissivity for all possible pairs of observation points you can form from the three existing measurement locations. See sheet Ex9.1a.

- **Exercise 9.1b**

Build a numerical groundwater model describing Thiem's problem. Use a 5 meter grid spacing (a) and assume that 195 metres is long enough so that the effect of pumping ceases to be seen in the level of the hydraulic head. Due to symmetry, it is enough to examine one fourth of a circle. The pumping cell is located in the upper left corner.

Equation (9.14) can be extended to two dimensions as

$$\frac{d^2H}{dx^2}+\frac{d^2H}{dy^2}=-\frac{R}{T} \tag{9.18}$$

where *x* and *y* are the location coordinates. Take a look at (9.15), discretize (9.18), and solve for the hydraulic head at a given location in the computation grid.

To build the numerical groundwater model you need to

- insert an expression for the hydraulic head in each cell in the calculation grid
- define correct boundary conditions on each side of the calculation grid

After you have created your model, transform the discharge of the pumping well into a sink term, and assign a transmissivity value that you have estimated using Thiem's equation.

1. Try to reproduce the drawdown data you were given in Exercise 9.1a. How close a match do you get? Discuss reasons why the match is not perfect.

Hint: While building the model, it may be sensible to turn the automatic calculation off. Then you do not need to wait for Excel to do calculations every time you edit a cell. To do this, go *Tools->Options->Calculation* and select *Manual Calculation*. When you want to calculate the sheet, hit F9 (or press *Tools->Options->Calculation Calc Now* button). When you are calculating your final results, calculate the sheet a few times until the numbers in the calculation grid do not change anymore. Then you know that the iteration has converged, 1000 iterations may not be quite sufficient.

- **Exercise 9.1c**

This is the most demanding of Section 9.1 exercises and relies heavily on the theory of non-linear optimisation presented in Section 4.4. The objective is to solve the inverse groundwater problem where values of hydraulic properties (e.g. transmissivities) are estimated using measurements of hydraulic heads. This is an inverse procedure to the forward problem of simulating hydraulic head distribution with known hydraulic property values, which explains the term inverse.

The inverse groundwater problem is generally ill-posed (see also Section 4.5), which means that a solution may not exist at all, the solution may not be unique, or the solution may not be stable. One might think that existence is guaranteed as the real values of the hydraulic properties should provide an accurate solution to the inverse problem. But as a result of measurement and model errors it is likely that no combination of hydraulic property values yields an exact match to measured hydraulic heads, and hence an exact solution to the inverse problem may not exist. This, however, is not a severe limitation as the real hydraulic property values provide at least an approximate solution to the problem.

Non-uniqueness and instability is an often-encountered problem in inverse groundwater modelling. Non-uniqueness refers to the situation where different combinations of hydraulic properties give similar hydraulic head values. It is impossible to know which combination represents the real situation and hence wrong hydraulic property values may be assigned to the groundwater model. Although different sets of hydraulic property values (which could be wrong) yield an equal fit to the available observations, they do not necessarily produce similar results in model applications outside of the observation period. See *Nonuniqueness* in Section 4.5. This of course is problematic and may lead to erroneous water level predictions.

In stable problems a small change in input values results in a small change in output. The forward groundwater problem is known to be stable, which means that small changes in hydraulic properties of the aquifer always leads to small changes in the simulated hydraulic head distribution. The inverse problem, however, is often unstable. The identified hydraulic properties can change significantly when the hydraulic head distribution changes only slightly.

For further reading on the inverse groundwater problem, the textbook by Sun (1994) is recommended.

This exercise comes with an Excel template (CMWRA_9_1.xls), where a simple groundwater flow model describing one-dimensional steady flow in a confined aquifer has been constructed. The aquifer domain has been split into ten inner nodes and two boundary nodes. The boundary condition at the right end is of constant head type, and at the left end the user can select either a constant head or a constant flux boundary condition. The aquifer has two zones which can have different transmissivity values. Within each zone transmissivity does not change from node to node.

In the upper left corner the user can change parameters that affect the optimisation procedure (*Optimisation Parameters*). These include the Marquardt coefficient λ (*Marquardt* λ), the perturbation value for numerically approximating derivatives (*Perturbation*), and the number of iterations to be carried out (*Iterations*).

Beside the optimisation parameters the template has parameters for determining the boundary conditions. In the *Head* cell the user specifies the prescribed head at the left end of the aquifer. And the *Flux* cell has the value of the prescribed flux at the left end of the aquifer. In this exercise you do not need to change these values.

In the *Model Fit* box the user supplies measured values for hydraulic heads in those nodes from which observed data are available. Measurements have been taken at two locations (one in each zone). In the Excel template these locations are marked with *Obs*1 and *Obs*2. The *Calculated* column contains calculated head values (i.e. model results) for the same locations. Cell O16 shows the sum of squared errors.

Below the groundwater model there are three buttons. The *Optimise* button is assigned to a macro that carries out the iteration. This macro copies the updated transmissivities (*Updated Parameters* in the template, cells C25 and D25) to the row 22 to be used in the groundwater model. Then Excel will calculate new head values according to the new, updated transmissivities (**Note! You need to have the Iteration box checked in Tools->Options->Calculation**). After the model has been run the macro will again copy the updated transmissivities to the row 22. This cycle is repeated as many times as specified in *Iterations* (cell C15).

Recover button becomes handy if the groundwater model collapses. This will happen if any of the transmissivities the model uses (cells C22, D22, C23, D23, C24, D24) get a zero or a negative value. Then you will see errors like #VALUE, #NULL or #DIV0. To recover the situation and to get rid of these errors hit the *Recover* button, which is assigned to a macro that will recreate the groundwater model. **Important: before using *Recover* button make sure that all transmissivities (cells C22, D22, C23, D23, C24, D24) have valid values!** *Optimise* macro checks that the updated transmissivities copied to the row 22 have positive values. If this is violated, optimisation will stop. If this happens, there may be a mistake in the Levenberg-Marquardt algorithm – or if not, the value of the Marquardt coefficient λ should be increased.

Change Boundary Condition button – as its name suggests – changes the type of the boundary condition at the left end of the aquifer. Text in cell E21 (*Head* or *Flux*) indicates which boundary condition is currently active. When the *Change Boundary Condition* button is pressed, the boundary condition parameters in cells F13 and F14 are also automatically updated, so the user does not need to touch them.

Construct in the Excel template the Levenberg-Marquardt algorithm for estimating transmissivities from the given hydraulic head data. Use matrix operations available in Excel when building the algorithm. In the template below the groundwater model there is space reserved for the matrices and vectors you will need. These spaces are set in such a manner that each new matrix created is a result of just one matrix operation (e.g. transpose, multiplication, inverse).

Start with the supplied value of 0.001 for the Marquardt coefficient λ, it probably will work fine in this exercise. But if the algorithm does not work (does not converge, or gives negative values for the transmissivity) – and you believe there is no mistake in the algorithm – you can try to increase this value e.g. to 0.01.

You need values for the gradient of the errors with respect to parameters T1 and T2 in constructing the sensitivity matrix (see (4.36)). Approximate the gradient using the difference quotient, which is defined as the function difference divided by the point difference. When function F depends on vector $\mathbf{x}$ the difference quotient is calculated by perturbing $\mathbf{x}$ by $\Delta\mathbf{x}$ and evaluating the function value both at $\mathbf{x}$ and $\mathbf{x} + \Delta\mathbf{x}$. The difference quotient is then given by

$$\frac{F(\mathbf{x}+\Delta\mathbf{x})}{\Delta\mathbf{x}}$$

Clearly when $\Delta\mathbf{x}$ becomes infinitesimally small the difference quotient equals the derivative of function *F*. When approximating the gradient of the errors with respect to parameters T1 and T2 use the supplied perturbation of 0.01 (*Perturbation*, cell C14). **Use *Recover* button after you have supplied cells C23, D23, C24 and D24 with the correct perturbation formulae that give positive values for transmissivities!**

Provide answers to the following tasks:

1. This exercise can be solved without the Levenberg-Marquardt optimisation algorithm. Select the prescribed head boundary condition. The observed hydraulic head at the first observation point (*Obs1*) is 7.78 m and at the second point (*Obs2*) the observed head value is 5.56 m. It is known that T1 = 10 $m^2\ d^{-1}$ and T2 = 40 $m^2\ d^{-1}$ is one solution to the problem. What is the sum of squared errors then? Demonstrate that the known solution is not unique.
2. Set the boundary condition to prescribed flux. The observed hydraulic head at the first measurement point (*Obs1*) is 7.60 m, and at the second point (*Obs2*) 5.75 m. It is known that both transmissivities T1 and T2 fall within a range from 5 $m^2\ d^{-1}$ to 100 $m^2\ d^{-1}$. Answer the following questions:
    a) Given the observations, what are the optimal transmissivities $T1^{opt}$ and $T2^{opt}$?
    b) What is the sum of squared errors with the optimal transmissivity values?
    c) What was your initial guess for the transmissivity values?
    d) How many iteration cycles were required to reach the optimal transmissivity values? And how many model runs were performed in those iteration cycles?

9.2 Macropore flow

THEORY

Sometimes we need to assume that soils are homogeneous although this assumption may be far from the reality. Soils often contain conduits that are much larger than the average pore size. These conduits – typically referred to as macropores – can be caused, for example, by plant roots, worms, or soil cracks.

When the hydraulic conductivity of soil outside of macropores is very small, the role of macropores will become very significant as they can deliver most of the water in the soil profile. Figure 9.9 shows a photograph of a clay profile where dye has been introduced to track the flow of water. It can clearly be seen that most of the water infiltrates down a small worm burrow.

Figure 9.9. Route of water in clayey soil traced using methylene blue dye. Photo: Visa Nuutinen.

Flow of fluid can be classified as either laminar or turbulent. When flow is laminar, the flow paths of the fluid particles do not cross each other. In turbulent flow the paths do cross and the motion of fluid particles can be described as irregular or chaotic (Figure 9.10).

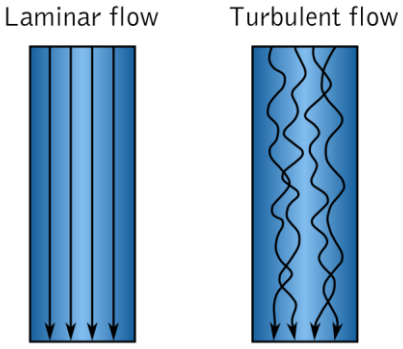

Figure 9.10. The difference between laminar and turbulent flow.

Macropores can be characterized as small tubes in the soil. Chen and Wagenet (1992) suggested that for describing laminar flow in macropores the Hagen-Poiseuille's equation be used, and for turbulent macropore flow the Manning equation be applied. They also assumed that the threshold value of the Reynolds number for determining the flow pattern is 3.0, below which the flow is laminar and above which it is turbulent.

The Reynolds number Re [-] is computed from

$$\mathrm{Re} = \frac{Ud}{\upsilon_k} \tag{9.19}$$

where U [m s^{-1}] is the average velocity in the tube, d [m] is the diameter of the tube, and υ_k [m^2 s^{-1}] is the kinematic viscosity of water. The Hagen-Poiseuille's equation is given as

$$U = \frac{gr^2}{8\upsilon_k}\frac{\Delta h}{\Delta z} \tag{9.20}$$

where g [ms^{-2}] it the acceleration of gravity, r [m] is the internal radius of the tube and $\Delta h/\Delta z$ [-] is the hydraulic gradient (where Δh is the change in the hydraulic head and Δz is the length of the tube).

The Manning equation can be written as

$$U = \frac{1}{n}\left(R^{\frac{2}{3}} S_p^{\frac{1}{2}} \right) \tag{9.21}$$

where n [sm$^{-1/3}$] is the coefficient of roughness, R [m] is the hydraulic radius, and S_p (or $\Delta h/\Delta z$) [-] is the slope of the energy line. We will assume that the tube is filled with water and hence the hydraulic radius R is equal to $r/2$.

EXERCISE 9.2

Objective

The objective of this exercise is to gain insight into the significance of macropores in subsurface water flow.

Task

The task is to calculate the radius of a single vertical macropore that conducts the same amount of water as a clay soil profile free of macropores. You can apply the unit hydraulic gradient ($\Delta h/\Delta z$=1) to both Hagen-Poiseuille and Manning equations. This means that the change in the hydraulic head is equal to the change in the elevation of a water particle moving down the macropore, i.e. the pressure of the water is assumed constant everywhere in the macropore.

Assume that the hydraulic conductivity of clay is K = 10^{-9} m/s. For the coefficient of roughness of the macropore use the value n = 0.13 (Chen & Wagenet 1992). Answer the following question:

1. What is the radius of a single macropore that conducts the same rate of water as a clay soil profile free of macropores and with an area of
   a) 0,5 m^2 ?
   b) 5 m^2 ?

See also Figure 9.11.

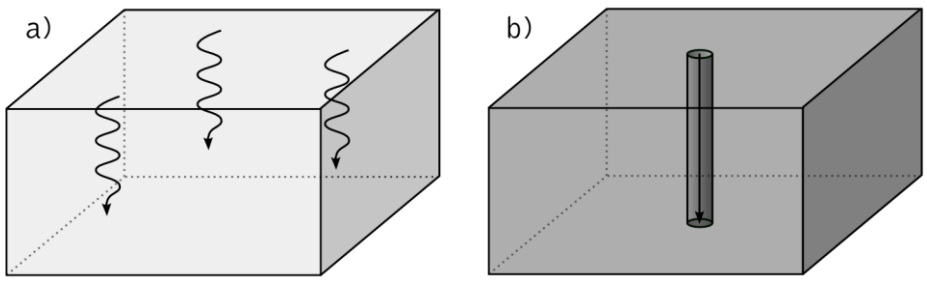

Figure 9.11. Water infiltrating through the clay profile (a) and flowing through the macropore (b).

References

Chen C. & Wagenet R.J. 1992. Simulation of water and chemicals in macropore soils Part 1. Representation of the equivalent macropore influence and its effect on soilwater flow. Journal of Hydrology,130, 105-126.

Sun, N-Z. 1994. Inverse Problems in Groundwater Modeling, Kluwer Academic Publishers, Dordrecht, the Netherlands.

Thiem, G. 1906. Hydrologic Methoden, J. M. Gebhardt, Leipzig, Germany.

10 SOIL WATER

T. Kokkonen, L. Stenberg

10.1 Soil water retention curve

THEORY

Water beneath the ground is referred to as groundwater when it exists below the groundwater table, and it is called soil water when it resides above the groundwater table in the unsaturated zone. The zone below the groundwater table is called the saturated zone as all voids between the soil grains are filled with water. Similarly, the zone between the groundwater table and the soil surface is called the unsaturated zone as the voids are only partly filled with water – the remainder being filled with air (Figure 10.1).

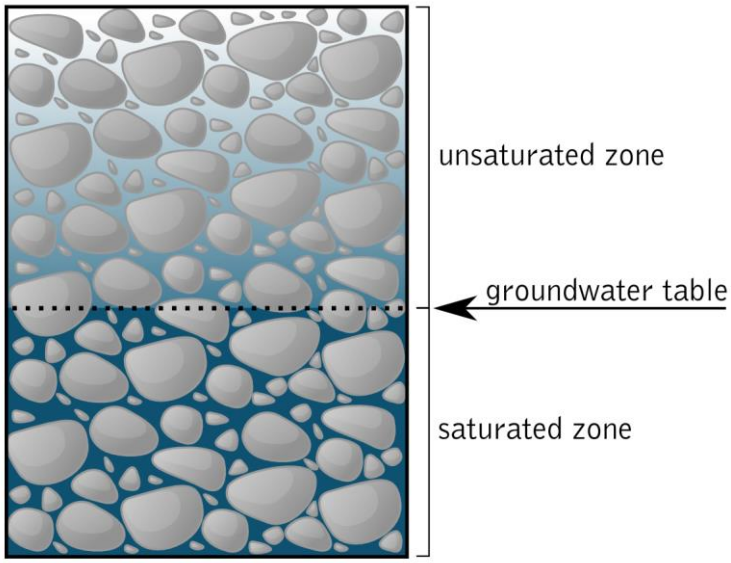

Figure 10.1. Schematic illustration of soil. Above the groundwater table the volumetric water content decreases (as the colour fades).

In the unsaturated zone the water is retained in the soil by capillary and molecular forces. As a result of these forces the pressure in soil water is smaller than the atmospheric pressure, which can easily be illustrated by taking a moist soil sample (or a sponge) in hand and observing that the retained water will not leak out. The atmospheric pressure is typically considered as the reference pressure in subsurface hydrology, and hence the pressure above the atmospheric pressure is positive, and similarly the pressure below the atmospheric level is negative. The negative pressure is also often called tension or suction. The pressure at the groundwater table is atmospheric, i.e. zero according to the above convention. In hydrostatic conditions (i.e. there is no flow of water in the soil profile) the suction also gives the distance from the groundwater table.

The relationship between the volumetric water content and the soil water pressure can be described with a soil water retention curve (soil moisture characteristic curve) that plots the soil water pressure as a function of the soil water content. Often the soil water pressure is given as a pF value, which is the base-10 logarithm of the pressure in centimetres in the water column h, i.e.

$$\mathrm{pF} = \log_{10}(-h) \tag{10.1}$$

Figure 10.2 plots two examples of soil water retention curves. When the pressure is zero, the corresponding soil water content equals the porosity of soil (note that due to the logarithmic scale the soil water retention curves in the figure start from a pressure value of -1 cm). When moving up from the groundwater table the water content in soils with finer grains (e.g. clays, Figure 10.2a) decreases much less than in coarse-grained soils (e.g. sand, Figure 10.2b). This is due to stronger capillary forces acting in fine-grained soils where the 'tubes' formed by interconnected voids in soil are narrower.

The soil water retention curve is measured by applying a successively higher suction (suction is defined as the negative of pressure) to a soil sample. The largest voids drain first at lower suction values, and when suction is increased smaller and smaller voids release the water stored in them. Recording the pairs of suction and the corresponding water content of the soil sample, the soil water retention curve can be constructed.

In hydrostatic conditions, i.e. when there is no movement of water in the soil profile, the soil water retention curve also determines the vertical water content distribution in a soil profile. Figure 10.3 shows the distribution between solid grains, water and air in a homogeneous soil profile under a hydrostatic condition with two different groundwater levels.

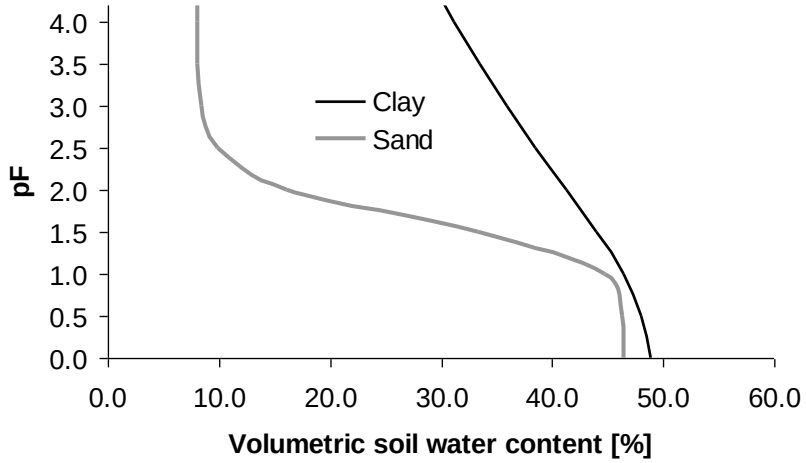

Figure 10.2. Soil water retention curves for clay (Kankaanranta, 1996) and sand (Jauhiainen, 2004).

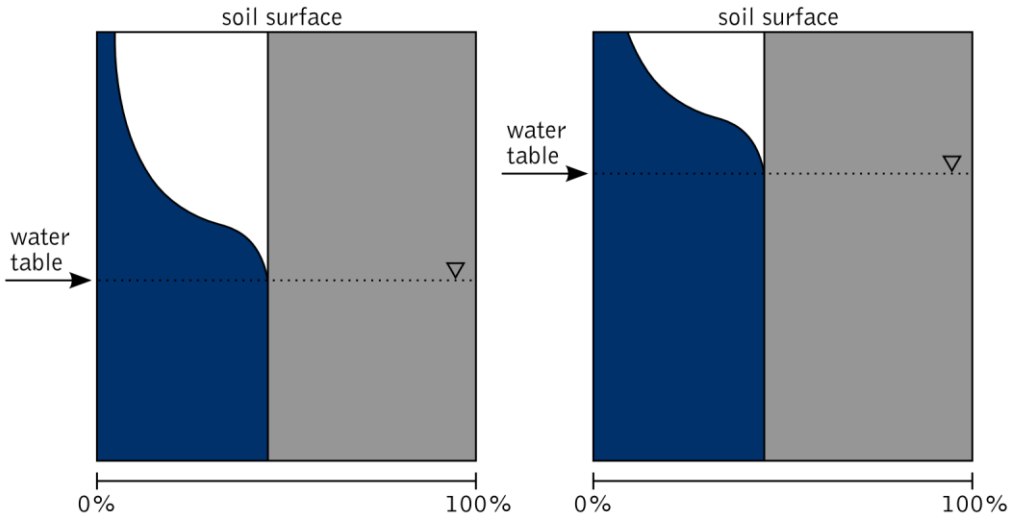

Figure 10.3. Distribution (volumetric) between solid grains (grey), water (blue), and air (white) in a homogeneous soil profile under a hydrostatic condition with two different groundwater levels.

EXERCISE 10.1

Objective

The objective is to understand the concept of the soil water retention curve and to learn how to use it.

Background

In workbook CMWRA_10_1.xls you have been given a series of suction – water content pairs for two soils. *Soil 1* represents sand and *soil 2* clay. The resulting soil water retention curves are plotted in a graph in sheet Ex10.1.

In this exercise we will assume that during times of observation the soil profile is always under hydrostatic conditions, i.e. there is no movement of water in the soil profile.

Task

- **Exercise 10.1a**

Consider a soil profile that is four metres thick and consists of *soil 1*. Initially the groundwater table is at a depth of one metre below the soil surface. At a later time the groundwater table has settled at a depth of two metres.

Answer the following questions:

1. How much water (in centimetres of water column) has been removed from the entire soil profile between the two observation times?
2. Why is the depth of water removed from the soil profile not equal to the change in the groundwater table?
3. How much water has been removed from the layer between the soil surface and one metre below the soil surface?
4. Plot in one graph the vertical distribution of soil water content for both groundwater levels.

- **Exercise 10.1b**

Consider two soil profiles that are both four metres thick. The first one consists of *soil 1* and the second consists of *soil 2*. The groundwater table is at a depth of two metres below the soil surface.

Answer the following questions:

1. Which of the two profiles retains more water?
2. How much water should be added to each of the two profiles to raise the groundwater table by one metre?

- **Exercise 10.1c**

Now consider a soil profile with two different soil layers (see Figure 10.4). The profile is four metres thick and the groundwater table is at a depth of three metres below the soil surface. The bottom layer is three metres thick and consists of *soil 1*. The top layer is one metre thick and consists of *soil 2*.

Answer the following questions:

1. How much water is stored in the entire profile?
2. How much water needs to be added to raise the groundwater table by one metre? Compare your result to Exercise 10.1b and explain why the required water amount here is different?
3. How much water needs to be added to raise the groundwater table by one metre if the bottom layer is of *soil 2* and the top layer of *soil 1*?
4. Plot in a single graph the vertical distribution of both soil water content and soil water pressure for the entire profile.

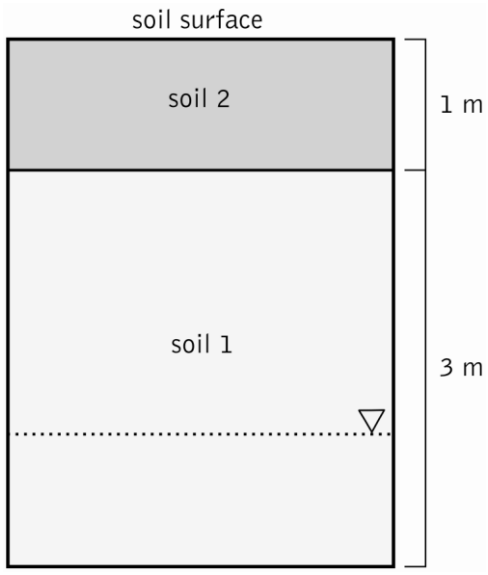

Figure 10.4. Soil profile in Exercise 10.1c.

References

Jauhiainen, M. 2004. Relationships of particle size distribution curve, soil water retention curve and unsaturated hydraulic conductivity and their implications on water balance of forested and agricultural hillslopes. Helsinki University of Technology, Water Resoures Publications, TKK-VTR-12, Espoo, 165 pp.

Kankaanranta, J. 1996. Water flow and nutrient leaching in clay soils (in Finnish: Veden virtaus ja ravinteiden huuhtoutuminen savimaassa). Master's Thesis, Helsinki University of Technology, Laboratory of Water Resources. 79 pp.

11 RESERVOIR WATER BUDGET

L. Stenberg, T. Kokkonen

11.1 Estimating lake inflow

THEORY

Lake inflow can be estimated using discharge measurements from streams flowing to the lake (when such data are available), or by formulating the water balance for the lake and estimating the inflow as the only unknown quantity while assuming all other water balance components to be known.

Water balance for a lake can be formulated as

$$Q_{IN} = Q_{OUT} + E - P + \frac{\Delta S}{\Delta t} \tag{11.1}$$

where Q_{IN} [mm d^{-1}] is the lake inflow, Q_{OUT} [mm d^{-1}] is the lake outflow, E [mm d^{-1}] is the evaporation from lake, P [mm d^{-1}] is the precipitation into the lake, ΔS [mm] is the change in the storage volume of the lake, and Δt [d] is the length of the observation period.

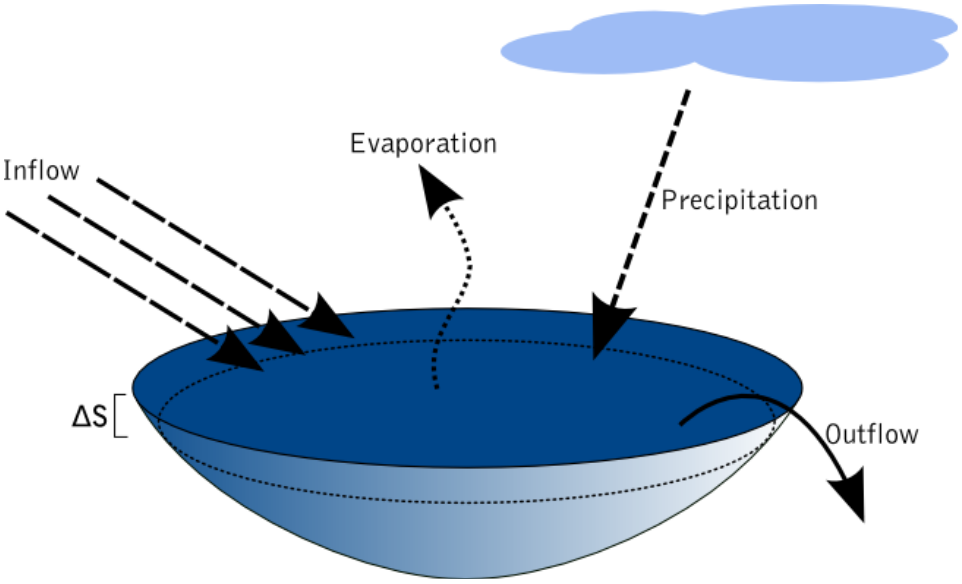

Figure 11.1. Incoming and outgoing water fluxes of a lake.

The change in the storage volume ΔS can be determined from the water level of the lake when the stage-volume relationship of a lake is known (Figure 11.2).

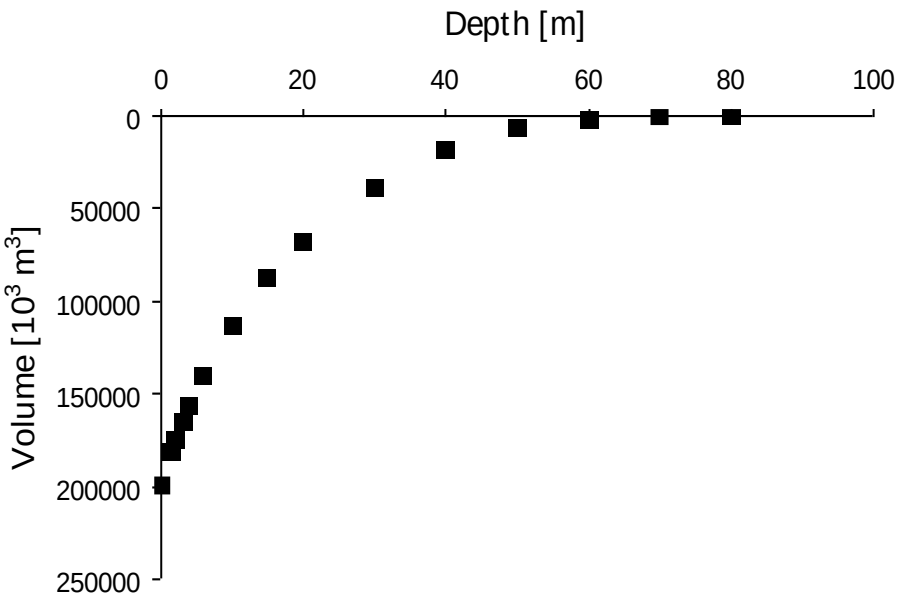

Figure 11.2. Lake depth – volume relationship for Lake Pääjärvi, Finland.

EXERCISE 11.1

Objective

The objective of this exercise is to learn to estimate lake inflow and to understand the processes that are involved in the water balance of a lake.

Meteorological and streamflow data

The discharge data and Class A evaporation data have been kindly provided by the Finnish Environment Institute and the precipitation data by the Finnish Meteorological Institute. It is not permitted to use this data for purposes other than solving this exercise.

Background

As in the lake evaporation exercise discussed in Section 6.1, the data for this exercise are also from Lake Pääjärvi. Lake Pääjärvi has five streams flowing to the lake, and the catchments of those five streams comprise ca. 84 % of the land area of the entire catchment of the lake (Figure 11.3).

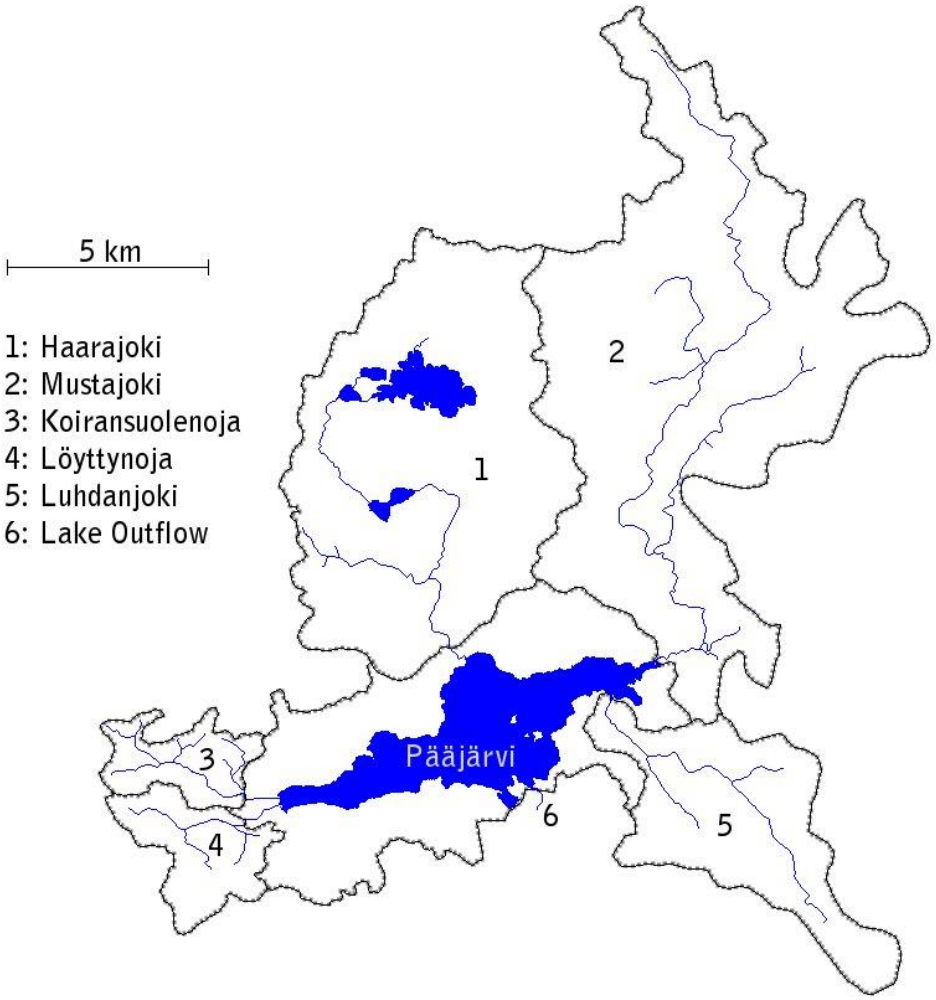

Figure 11.3. Lake Pääjärvi and its catchment area.

Complete discharge data from all streams flowing to a lake are rarely available. Discharge data may be completely missing, or the fraction of the entire lake catchment covered by the catchments of gauged streams may be too small. In such a case the mass balance approach presented in (11.1) can be applied to estimate lake inflow – given that lake outflow data, meteorological observations, and lake water level measurements are available, along with the relationship between the water level and the volume of the lake.

For this exercise an estimate of lake evaporation is provided. Due to lack of radiation data for the time period considered here, the evaporation estimate is not computed by the combination equation discussed in Section 6.1 but it is based on Class A pan measurements from Vestola station located on the ground 9 km from Lake Pääjärvi. Between the years 1971 and 1974 lake evaporation measurements were conducted using a GGI-3000 evaporation pan, and the relationship between Class A and GGI-3000 measurements in those years was used to adjust Class A measurements throughout the study period. Note that Class A measurements are not available from winter (October – April). Due to the following two reasons it will be assumed here that missing winter data have only a minor impact on the estimated lake inflow. First, evaporation form the lake surface is low during the winter time. This is particularly true when the lake has an ice-cover. And second, the fraction of the lake area from the catchment area of the lake is relatively small (6.4%), and therefore lake evaporation is small compared to the inflow to the lake.

Task

Calculate inflow to Lake Pääjärvi in two different ways using the data provided in the Excel workbook CMWRA_11_1.xls (sheet Ex11.1). First, use streamflow data to estimate the lake inflow for the time period from 1981 to 1993. For the first years of the period you have data from four out of the five sub catchments (Figure 11.3) covering 73% of the total catchment land area, and from 1987 onwards you have data from three sub catchments (70% of the total catchment land area). The areas of the sub catchments are listed in the top left corner of sheet Ex11.1. The cells containing the areas have names (*A_musta, A_haara, A_koira, A_loytty*, and *A_catchment*) which you can use.

Second, apply the water balance equation (11.1) to estimate the lake inflow. To calculate the change in storage volume you can use the following polynomial function to describe the relationship between the lake water level W [m] and the storage volume S [m^3]

$$S = aW^2 + bW + c \tag{11.2}$$

where a, b, and c are coefficients whose values are given in Excel sheet Ex11.1. In sheet Ex11.1 you will also find discharge data for lake outflow (river Teuronjoki) and the estimated lake evaporation.

On the right hand side of the sheet there are tables and graphs summarising the results. Based on your calculations, answer the following questions:

1. Usually when precipitation increases, runoff would also increase. Compare the time periods from Nov 1981 to Oct 1982 and from Nov 1982 to Oct 1983. What do you observe? Discuss the reasons that might explain your observation. What kind of climatic conditions could have prevailed in those two time periods? What kind of additional data would you like to have to verify your explanation?
2. In March 1988 there is one day when the lake inflow estimated using the water balance equation is negative (smaller than –4 mm d^{-1}). What could cause this unreasonable value?

12 SPATIAL HYDROLOGY

T. Kokkonen, A. Jolma

Note Jan 30, 2020: Goinformatica GIS software (Section 2.2.) is no longer included in the distribution package of this exercise book. Exercise 12.1 can be solved using other GIS software. Note that the convention for flow direction coding (Figure 12.1) can differ between different GIS software.

12.1 Catchment and stream network delineation

THEORY

Landscape topography is a major hydrological driver, which gives the basis for identifying stream network and catchment structures using a digital elevation model (DEM). A DEM contains gridded data about the soil surface elevation above a reference level. In the identification of hydrologic pathways, one of the simplest and most commonly used methods assumes that the direction of flowing water coincides with the direction of the steepest descent. This method where one of the eight directions to the neighbouring grid cells is selected is commonly known as the D8 method (O'Callaghan and Mark, 1984, Figure 12.1).

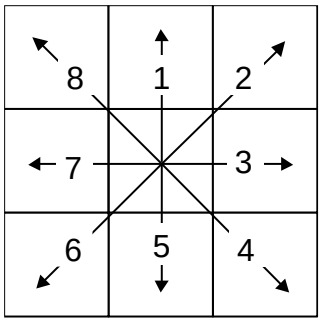

Figure 12.1. Possible flow directions from a grid cell in the D8 method.

Based on the flow directions (Figure 12.2a and Figure 12.2b) the catchment of a location (grid cell) can be identified (Figure 12.2d). Catchment is defined as the area whose waters drain through the given grid cell. Also, once flow directions are available flow accumulation values depicting the size of the upslope area can be computed for each grid cell (Figure 12.2c). Flow accumulation values, which indicate how many grid cells drain through a given cell, can be used in stream network delineation. Following the assumption that places with large upslope areas are likely to form streams, a stream network can be delineated simply by applying a threshold value to the flow accumulation grid. All grid cells having a greater flow accumulation value than the threshold value are assumed to be stream cells (Figure 12.2c).

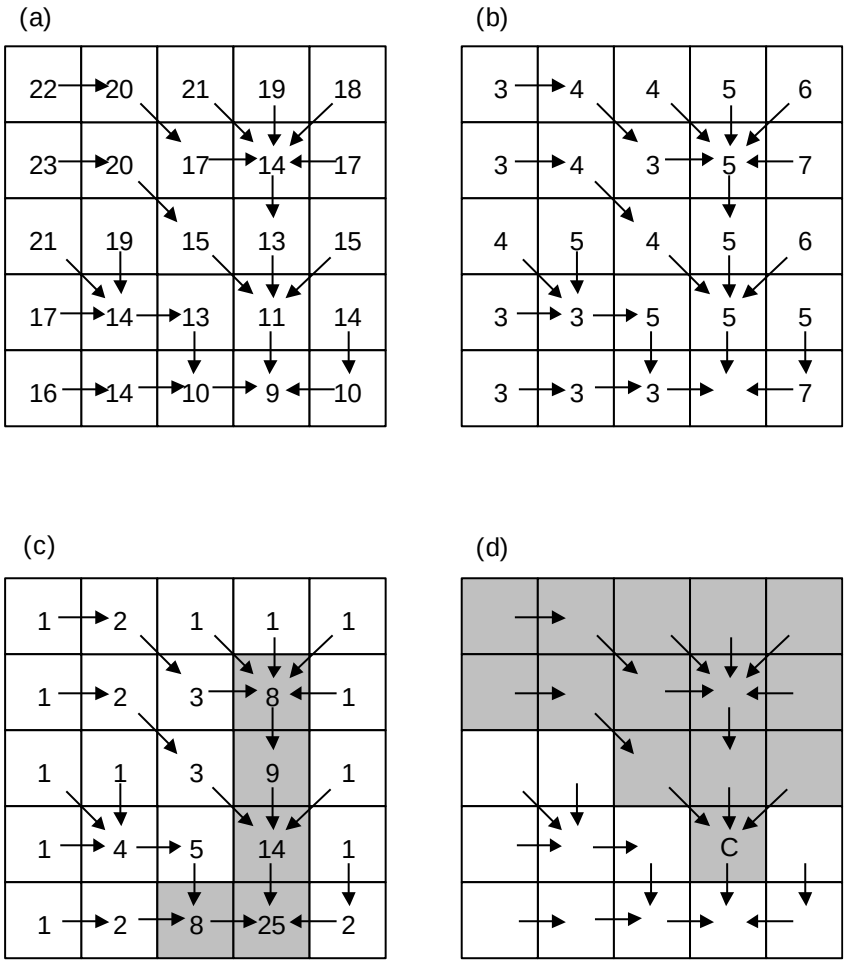

Figure 12.2. Derivation of flow directions in the D8 method. Elevations of grid cells (a); flow direction coding using the convention presented in Figure 1 (b); flow accumulation values and prescribed stream cells (grey) using a threshold value > 7 are also shown (c); catchment (grey) of the grid cell indicated by 'C' (d).

Digital elevation models typically contain flat areas and local depressions (pits), which complicates routing of water through the landscape. Water can be forced to flow through flat areas by iteratively directing the flow to a neighbour that has a resolved flow direction (i.e. to a cell that is neither a flat nor a pit). Pits are typically handled by spilling the water to the lowest grid cell surrounding the pit. Interested readers are referred e.g. to Martz and Garbrecht (1998) for further reading on the subject.

While the D8 method is widely applied, it does suffer from imprecision arising from the coarse discretization of the flow direction into only eight possible values. Interested readers are referred to Fairfield and Leymarie (1991), Quinn et al. (1991), Costa-Cabral and Burges (1994), and Tarboton (1997) for suggestions on how to improve the D8 method.

References

Costa-Cabral, M., and Burges, S. J. 1994. Digital elevation model networks (DEMON): A model of flow over hillslopes for computation of contributing and dispersal areas. Water Resources Research, 30(6), 1681–1692.

Fairfield, J., and Leymarie, P. 1991. Drainage networks from grid digital elevation models. Water Resources Research, 27, 709–717.

Martz, L.W., and Garbrecht, J. 1998. The treatment of flat areas and depressions in automated drainage analysis of raster digital elevation models. Hydrological Processes, 12, 843-855.

Quinn, P., Beven, K., Chevallier, P., and Planchon, O. 1991. The prediction of hillslope flow paths for distributed hydrological modeling using digital terrain models. Hydrological Processes, 5, 59–80.

EXERCISE 12.1

Objective

The objective of this exercise is to obtain an appreciation of how watersheds and stream networks can be delineated using digital elevation data.

Data

The DEM available for this exercise has been kindly provided by the National Land Survey of Finland. It is not permissible to use the elevation data for purposes other than solving this exercise.

Vector data depicting the location of the River Vantaanjoki has been compiled from the CORINE Land Cover 2000 data (http://www.environment.fi/syke/clc2000).

Task

You have digital elevation data for the Vantaanjoki basin area in southern Finland as a grid data set called *vantaa_dem.tif*. After you have opened the grid data set remember to load it to RAM (see the important note in *Opening data* in Section 2.2). Use the terrain analysis functions available in Geoinformatica to delineate the watershed above the Myllymäki gauging station in the River Vantaanjoki. The coordinates of the gauging station are X = 3382122 and Y = 6689131. The reported drainage area above the Myllymäki gauging station is 1230 km^2. Answer the following questions:

1. How does the drainage area of the watershed that is delineated using the DEM compare with the reported value?
2. What is the fraction of areas that are flat according to the DEM in the Myllymäki watershed? The flow direction coding for a flat area in Geoinformatica is –1 (0 denotes a pit).

You also have the map location of the River Vantaanjoki as a vector data set called *vantaa*. Load the data set into Geoinformatica to be used as a reference when delineating the stream network. Answer the following questions:

3. In your opinion, what is an appropriate flow accumulation threshold value for the delineation of the stream network? All grid cells having a greater flow accumulation value than the threshold are interpreted as stream cells.
4. In what type of regions does the delineated stream network differ most from the real location of the river?

Hints

You will find terrain analysis (*Geo::Raster::TerrainAnalysis*) and zonal functions (*Geo::Raster::Zonal*) of Geoinformatica useful in this exercise. Find descriptions for these functions from Perl modules documentation accessible through the Geoinformatica start menu.

Note that due to inaccuracies in the elevation data, as well as in the flow direction computation algorithm, the location of delineated streams typically differs from the mapped location of streams in nature. Consider this when delineating the Myllymäki watershed.