GEO4-4420

Stochastic Hydrology

Prof. dr. Marc F.P. Bierkens
Department of Physical Geography
Utrecht University

Contents

Chapter 1: Introduction

1.1 Why stochastic hydrology?

The term "stochastic" derives from the Greek word στοχαστιχηζ which is translated with "a person who forecasts a future event in the sense of aiming at the truth", that is a seer or soothsayer. So, "stochastic" refers to predicting the future. In the modern sense "stochastic" in stochastic methods refers to the random element incorporated in these methods. Stochastic methods thus aim at predicting the value of some variable at non-observed times or at non-observed locations, while also stating how uncertain we are when making these predictions.

But why should we care so much about the uncertainty associated with our predictions? The following example (Figure 1.1) shows a time series of observed water table elevations in a piezometer and the outcome of a groundwater model at this location. Also plotted are the differences (residuals) between the data and the model results. We can observe two features. First, the model time series seems to vary more smoothly then the observations. Secondly, there are noisy differences between model results and observations. These differences, which are called residuals, have among others the following causes:

- *observation errors*. Is it rarely possible to observe a hydrological variable without error. Often, external factors influence an observation, such as temperature and air pressure variations during observation of water levels;
- *errors in boundary conditions, initial conditions and input*. Hydrological models only describe part of reality, for example groundwater flow in a limited region. At the boundaries of the model values of the hydrological variables (such groundwater heads or fluxes) have to be prescribed. These boundary values cannot be observed everywhere, so there is likely to be some error involved. Also, if a model describes the variation of a hydrological system in time, then the hydrological variables at time step zero must be known as it determines how the system will be evolve in later time steps. Again, the initial values of all the hydrological variables at all locations are not exactly known and are estimated with error. Finally, hydrological models are driven by inputs such as rainfall and evaporation. Observing rainfall and evaporation for larger areas is very cumbersome and will usually be done with considerable error;
- *unknown heterogeneity and parameters*. Properties of the land surface and subsurface are highly heterogeneous. Parameters of hydrological systems such as surface roughness, hydraulic conductivity and vegetation properties are therefore highly variable in space and often also in time. Even if we were able to observe these parameters without error, we cannot possibly measure them everywhere. In many hydrological models parameters are assumed homogeneous, i.e. represented by a single value for the entire (or part of the) model region. Even if models take account of the heterogeneity of parameters, this heterogeneity is usually represented by some interpolated map from a few locations where the parameters have been observed. Obviously, these imperfect representations of parameters lead to errors in model results;

- *scale discrepancy.* Many hydrological models consist of numerical approximations of solutions to partial differential equations using either finite element or finite difference methods. Output of these models can at best be interpreted as average values for elements or model blocks. The outputs thus ignore the within element or within block variation of hydrological variables. So, when compared to observations that represent averages for much smaller volumes (virtually points), there is discrepancy in scale that will yield differences between observations and model outcomes;
- *model or system errors.* All models are simplified versions of reality. They cannot contain all the intricate mechanisms and interactions that operate in natural systems. For instance, saturated groundwater flow is described by Darcy's Law, while in reality it is not valid in case of strongly varying velocities, in areas of partly non-laminar flow (e.g. faults) or in areas of very low permeability and high concentrations of solvents. Another example is when a surface water model uses a kinematic wave approximation of surface water flow, while in reality subtle slope gradients in surface water levels dominate the flow. In such cases, the physics of reality differ from that of the model. This will cause an additional error in model results.

In conclusion, apart from the observation errors, the discrepancy between observations and model outcomes are caused by various error sources in our modelling process.

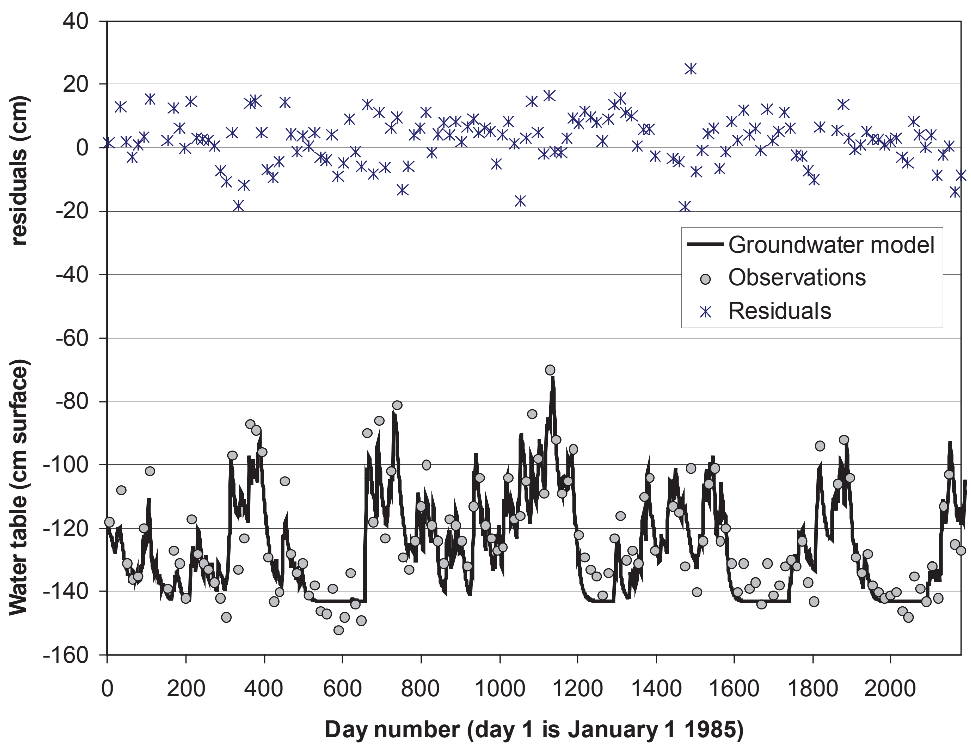

Figure 1.1 Observed water table depths and water table depths predicted with a groundwater model at the same location. Also shown are the residuals: the differences between model outcome and observations.

There are two distinct ways of dealing with errors in hydrological model outcomes:

Deterministic hydrology. In deterministic hydrology one is usually aware of these errors. They are taken into account, often in a primitive way, during calibration of models. During this phase of the modelling process one tries to find the parameter values of the model (e.g. surface roughness or hydraulic conductivity) such that the magnitude of the residuals is minimized. After calibration of the model, the errors are not explicitly taken into account while performing further calculations with the model. Errors in model outcomes are thus ignored.

Stochastic Hydrology. Stochastic hydrology not only tries to use models for predicting hydrological variables, but also tries to quantify the errors in model outcomes. Of course, in practice we do not know the exact values of the errors of our model predictions; if we knew them, we could correct our model outcomes for them and be totally accurate. What we often do know, usually from the few measurements that we did take, is some probability distribution of the errors. We will define the probability distribution more precisely in the next chapters. Here it suffices to know that a probability distribution tells one how likely it is that an error has a certain value.

To make this difference more clear, Figure 1.2 is shown. Consider some hydrological variable z, say soil moisture content, whose value is calculated (at some location and at some time) by a unsaturated zone model. The model output is denoted as $\breve{z}$. We then consider the error $e = \breve{z} - z$. Because we do not know it exactly we consider it as a so called *random variable* (chapter 3) E (notice the use of capitals for random variables) whose exact value we do not know but of which we do know the probability distribution. So in case of deterministic hydrology modelling efforts would only yield $\breve{z}$ (upper figure of Figure 1.2a), while stochastic hydrology would yield both $\breve{z}$ and the probability distribution of the (random) error E (lower figure of Figure 1.2a).

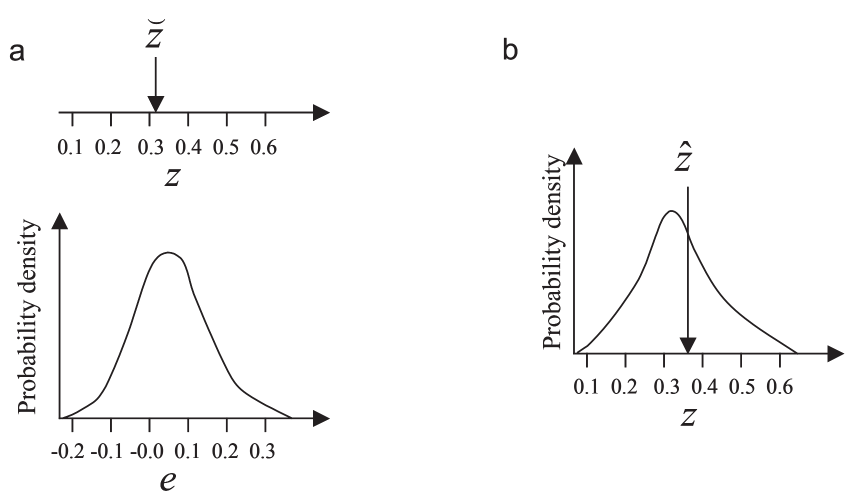

Figure 1.2 Stochastic Hydrology is about combining deterministic model outcomes with a probability distribution of the errors (Figure 1.2a), or alternatively, considering the hydrological variable as random and determining its probability distribution and some "best prediction"(Figure 1.2b).

Most of the methods used in stochastic hydrology do not consider errors in model outcomes explicitly. Instead it is assumed that the hydrological variable z itself is a random variable Z. This means that we consider the hydrological variable (e.g. soil moisture) as one for which we cannot know the exact value, but for which we can calculate the probability distribution (see Figure 1.2b). The probability distribution of Figure 1.2b thus tells us that although we do not know the soil moisture content exactly, we do know that it is more likely to be around 0.3 then around 0.2 or 0.5. Models that provide probability distributions of target variables instead of single values are called *stochastic models*. Based on the probability distribution it is usually possible to obtain a so called *best prediction* $\hat{z}$, which is the one for which the errors are smallest on average. Incidentally, the value of the best prediction does not have to be the same as the deterministic model outcome $\breve{z}$.

Box 1. Stochastic models and physics
A widespread misconception about deterministic and stochastic models is that the former use physical laws (such mass and momentum conservation), while the latter are largely empirical and based entirely on data-analysis. This of course is not true. Deterministic models can be either physically based (e.g. a model based on Saint-Venant equations for flood routing) and empirical (e.g. a rating curve used as a deterministic model for predicting sediment loads from water levels). Conversely, any physically based model becomes a stochastic model once its inputs, parameters or outputs are treated as random.

There are a number of clear advantages in taking the uncertainty in model results into account, i.e. using stochastic instead of deterministic models.

- The example of Figure 1.1 shows that model outcomes often give a much smoother picture of reality. This is because models are often based on an idealized representation of reality with simple processes and homogenous parameters. However, reality is usually messy and rugged. This may be a problem when interest is focussed on extreme values: deterministic models typically underestimate the probability of occurrence of extremes, which is rather unfortunate when predicting for instance river stages for dam building. Stochastic models can be used with a technique called "stochastic simulation" (see chapters hereafter) which is able to produce images of reality that are rugged enough to get the extreme statistics right.
- As stated above, the value of the best prediction $\hat{z}$ does not have to be the same as the deterministic model outcome $\breve{z}$. This is particularly the case when the relation between model input (e.g. rainfall, evaporation) or model parameters (e.g. hydraulic conductivity, manning coefficient) and model output is non-linear (this is the case in almost all hydrological models) and our deterministic assessment of model inputs and parameters is not error free (also almost always the case). In this case, stochastic models are able to provide the best prediction using the probability distribution of model outcomes, while deterministic models cannot and are therefore less accurate.
- If we look closely at the residuals in Figure 1 it can be seen that they are correlated in time: a positive residual is more likely to be followed by another positive residual and vice versa. This correlation, if significant, means that there is still some information present in the residual time series. This information can be used to improve model

predictions between observation times, for instance by using time series modelling (chapter 5) or geostatistics (chapter 6). This will yield better predictions than the deterministic model alone. Also, it turns out that if the residuals are correlated, calibration of deterministic models (which assume no correlation between residuals) yield less accurate or even biased (with systematic errors) calibration results when compared with stochastic models that do take account of the correlation of residuals (te Stroet, 1995).

- By explicitly accounting for the uncertainty in our prediction we may in fact be able to make better decisions. A classical example is remediation of polluted soil, where stochastic methods can be used to estimate the probability distribution of pollutant concentration at some non-visited location. Given a critical threshold above which regulation states that remediation is necessary, it is possible to calculate the probability of a false positive decision (we decide to remediate, while in reality the concentration is below the threshold) and that of a false negative (we decide not to remediate while in reality the concentration is above the threshold). Given these probabilities and the associated costs (of remediation and health risk) it is then possible for each location to decide whether to remediate such that the total costs and health risk are minimised.
- There are abundant stochastic methods where a relation is established between the uncertainty in model outcomes and the number of observations in time and space used to either parameterize or calibrate the model. If such a relation exists, it can be used for monitoring network design. For example, in groundwater exploration wells are drilled to perform pumping tests for the estimation of transmissivities and to observe hydraulic heads. The transmissivity observations can be used to make an initial map of transmissivity used in the groundwater model. This initial map can subsequently be updated by calibrating the groundwater model to head observations in the wells. Certain stochastic methods are able to quantify the uncertainty in groundwater head predicted by the model in relation to the number of wells drilled, their location and how often they have been observed. These stochastic methods can therefore be used to perform monitoring network optimization: finding the optimal well locations and observation times to minimise uncertainty in model predictions.
- The last reason why stochastic methods are advantageous over deterministic methods is related to the previous one. Stochastic methods enable us to relate the uncertainty in model outcomes to different sources of uncertainty (errors) in input variables, parameters and boundary conditions. Therefore, using stochastic analysis we also know which (error) source contributes the most to the uncertainty in model outcomes, which source comes second etc. If our resources are limited, stochastic hydrology thus can guide us where to spend our money (how many observations for which variable or parameter) to achieve maximum uncertainty reduction at minimum cost. An excellent book on this view on uncertainty is written by Heuvelink (1998).

1.2 Scope and content of these lecture notes

These notes aim at presenting an overview of the field of stochastic hydrology at an introductory level. This means that a wide range of topics and methods will be treated, while each topic and method is only treated at a basic level. So, the book is meant as an introduction to the field while showing its breadth, rather than providing an in depth treatise. References are given to more advanced texts and papers for each subject. The book thus aims at teaching the basics to hydrologists who are seeking to apply stochastic methods. It can be used for a one-semester course at third year undergraduate or first year graduate level.

The lecture notes treat basic topics that should be the core of any course on stochastic hydrology. These topics are: descriptive statistics; probability and random variables; hydrological statistics and extremes; random functions; time series analysis; geostatistics; forward stochastic modelling; state prediction and data-assimilation. A number of more advanced topics that could constitute enough material for a second course are not treated. These are, among others: sampling and monitoring; inverse estimation; ordinary stochastic differential equations; point processes; upscaling and downscaling methods, uncertainty and decision making. During the course these advanced topics will be shortly introduced during the lectures. Students are required to study one of these topics from exemplary papers and write a research proposal about it.

1.3 Some useful definitions for the following chapters

1.3.1 Description of a model according to system's theory

Many methods in stochastic hydrology are best understood by looking at a hydrological model from the viewpoint of system's theory. What follows here is how a model is defined in system's theory, as well as definitions for state variables, input variables, parameters and constants.

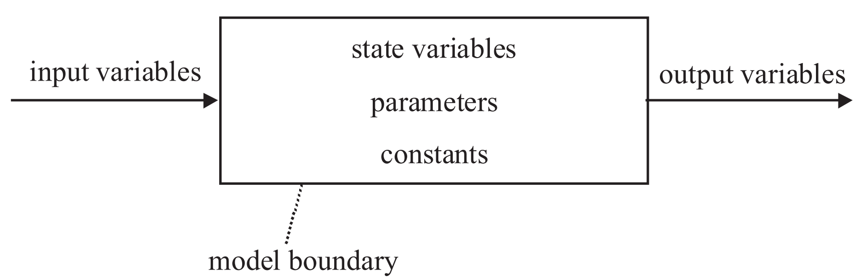

Figure 1.3 Model and model properties according to system's theory

Figure 1.3 shows a schematic representation of a model as used in system's theory. A *model* is a simplified representation of part of reality. The *model boundary* separates the part of reality described by the model from the rest of reality. Everything that is to know about the part of reality described by the model at a certain time is contained in the *state variables*. These are variables because their values can change both in space and time. The variation of the state variables is caused by the variation of one or more *input variables*. Input variables are always observed and originate from outside the model