# Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches Additional information

## Internal report

D. Koutsoyiannis

Department of Water Resources, National Technical University of Athens, Greece
Visiting scholar, Hydrologic Research Center, San Diego, CA., USA

A. Efstratiadis

Department of Water Resources, National Technical University of Athens, Greece

K. P. Georgakakos

Hydrologic Research Center, San Diego, CA., USA

**Abstract** This report contains additional information for the paper *Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches*. Specifically, it includes a summary of basic notions from the theory of probability and statistics.

## Probability space

Notation: $(\Omega, \Sigma, P)$

Probability space is the foundation of modern probability, according to Kolmogorov's axiomatic approach and the notion of *measure*, which maps sets onto numbers. In this, $\Omega$ is a non-empty set, sometimes called the *sample space* whose elements $\omega$ are known as *outcomes* or *states*; $\Sigma$ is a set known as *σ-algebra* whose elements $E$ are subsets of $\Omega$, known as *events* ($\Omega$ is also a member of $\Sigma$, called the *certain event*); and $P$ is probability described below.

## Probability

Notation: $P$

Probability is a function that maps events to real numbers, assigning each event $E$ (member of $\Sigma$) a number between 0 and 1. It must satisfy certain conditions, namely $P(\Omega) = 1$ and it must be a measure (meaning that $P(\ )$ should be a non negative number and have the additivity property, according to which the measure of the union of disjoint events $E_i$ is equal to the sum of the measures of all $E_i$ ).

## Random variable

Notation: $X$, $B$, $L$, $U$, … (upper case letters)

A random variable is a function that maps outcomes to numbers, i.e. quantifies the sample space $\Omega$. A random variable is not a number but a function. Values that a random variable may take in a random experiment, else known as *realizations* of the variables are numbers, usually denoted by lower case letters such as $x$, $b$, $l$, $u$, … Any event can be conveniently represented in terms of a random variable. For example, if $X$ represents the rainfall depth expressed in millimetres for a given rainfall episode (in this case $\Omega$ is the set of all possible rainfall depths) then $\{X \le 1\}$ represents an *event* in the probability notion (a subset of $\Omega$ and a member of $\Sigma$ – not to be confused with a physical event or episode) and has a probability $P\{X \le 1\}$. If $x$ is a realization of $X$ then $x \le 1$ is not an event but a relationship between the two numbers $x$ and 1, which can be either true or false. In this respect it has no meaning to write $P\{x \le 1\}$. Furthermore, if we consider the two variables $X$ and $L$ it has meaning to write $P\{X \ge L\}$ (i.e. $\{X \ge L\}$ represents an event) but there is no meaning in the expression $P\{x \ge l\}$.

## Distribution function

Notation: $F_X(x)$

By definition $F_X(x) := P\{X \le x\}$, i.e. it maps numbers ($x$ values) to numbers. The random variable to which this function refers is not an argument of the function; it is usually denoted as a subscript of $F$ (or even omitted if there is no risk of confusion). Typically $F_X(x)$ has some mathematical expression depending on some parameters $\beta_i$. These parameters are numbers.

## Sample of a random variable $X$

Notation: Classical statistics: $X_1, \ldots, X_n$; Modification: $\mathbf{X}_{0,n} = [X_0, \ldots, X_{1-n}]'$

In classical statistics a *sample* of $X$ of length $n$ is a sequence of independent identically distributed variables each having a distribution function identical to that of $X$. Each one may be viewed as representing the outcome of a random experiment. If we perform the experiment $n$ times, we obtain the $n$ numbers $x_0, \ldots, x_{1-n}$ which we call *observations*. Thus the sample is composed of random variables (functions) whereas observations are numbers.

In this study the following modifications of classical statistics have been made (1) the index is associated with time; (2) the arrangement of the sample members and observations are from the latest ($X_0$ and $x_0$, referring to the present) to the earliest ($X_{1-n}$ and $x_{1-n}$); (3) this arrangement is meant as a vector rather than a sequence; and (4) the random variables are stochastically dependent on each other rather than independent.

## Point estimator of a parameter $\beta$

Notation: $B := g_B(\mathbf{X}_0)$

Given any parameter $\beta$ related to the distribution function of a random variable $X$, a point estimator of $\beta$ is a random variable $B := g_B(\mathbf{X}_0)$ for some appropriate function $g_B(\ )$. $B$ is also called a *statistic*, as it is a function of the sample. For example, if $X$ represents the annual rainfall as a random variable, $\beta$ is the mean annual rainfall, which is a parameter (number – but not observable) and $X_i$ is the annual rainfall at year $i$ (random variable, observable), then $g_B(\mathbf{X}_{0,n}) = (X_0 + \ldots + X_{1-n}) / n$ is the sample mean (point estimator).

## Point estimate of a parameter $\beta$

Notation: $b = g_B(\mathbf{x}_{0,n})$

A point estimate of $\beta$ is a realization of $B$, i.e. the quantity $b = g_B(\mathbf{x}_{0,n})$. In the mean rainfall example, $g_B(\mathbf{x}_{0,n}) = (x_0 + \ldots + x_{1-n}) / n$ is the *observed* sample mean. In summary, one should distinguish among the three quantities: (1) a true parameter $\beta$ which is a number usually unknown in real world problems; (2) an estimate $b$ of this parameter, which is again a number, but known, calculated from the available observations; and (3) an estimator $B$ which is not a number but a random variable.

## Prediction limits of a random variable

Notation: $\lambda$ and $\upsilon$ in $P\{\lambda < X < \upsilon\} = \alpha$

If the variable $X$ lies in the interval $(\lambda, \upsilon)$ with probability $\alpha$, then the numbers $\lambda, \upsilon$ are called prediction limits of $X$ for *confidence coefficient* $\alpha$ and the tolerance interval $(\lambda, \upsilon)$ is called *prediction interval* of $X$ or *confidence interval of the prediction* of $X$.

## Confidence limits of a parameter

Notation: $L$ and $U$ in $P\{L < \beta < U\} = \alpha$

If the random variables $L$ and $U$ (standing for lower and upper, respectively) are related to the (unknown) parameter $\beta$ with the equation shown in the left, then they are called *confidence limits* of $\beta$ for confidence coefficient $\alpha$. Both $L$ and $U$ are functions of the sample (statistics), i.e. $U := g_U(\mathbf{X}_{0,n})$ and $L := g_L(\mathbf{X}_{0,n})$ for some appropriate functions $g_L(\ )$ and $g_U(\ )$. The random interval $(L, U)$, which brackets the parameter from both sides, is called the *confidence interval of the estimation* of $\beta$ and is an *interval estimator* of the parameter $\beta$. It should be stressed that while the prediction limits are numbers, the confidence limits of estimation are random variables, whose sample realizations $l$ and $u$ form the *interval estimate* of the parameter $\beta$.