European Water 83/84: 21-32, 2023.
© 2023 *E.W. Publications*

Estimation of the information content of spatially averaged precipitation data in lumped monthly hydrological simulation

E. Findanis* and A. Loukas

Laboratory of Hydraulic Works and Environmental Management, Department of Transportation and Hydraulic Engineering, School of Rural and Surveying Engineering, Aristotle University of Thessaloniki
* *e-mail: findanis@topo.auth.gr*

Abstract: Lumped runoff-rainfall models require a single time series of precipitation to simulate the runoff at the outlet of a watershed. When two or more precipitation gauge stations are available at different locations of the basin, their observations must be integrated into spatial average time series. Application of various methods of spatially averaged precipitation affects the estimation of areal precipitation and the performance of lumped models. Hence, to evaluate the suitability of each method, a selected model must be calibrated and validated. This process can be avoided when the information content of the spatially averaged precipitation is evaluated before the simulation. This is because, information decreases while estimating areal precipitation, since the initial set of time series is compressed into a single areal time series. The more information is retained after the act of spatially averaging, the more suitable the spatial distribution is. Estimates of the information content of a set of time series can result from Claude Shannon's Theory of Information. Moreover, unique spatial distributions may occur by minimizing the information drop that arises while estimating areal precipitation. In the present paper, the suitability of different averaging methods is evaluated by calibrating and validating five lumped models and comparing these results to the computed information drop. Furthermore, the informational content of models' input data is computed, and its relationship with the modeling performance is studied.

Key words: Lumped models; precipitation spatially averaging; calibration of hydrological models; theory of information

1. INTRODUCTION

Precipitation data at different gauge stations located in a watershed can be used to estimate spatially averaged areal precipitation time series, P, which could be used as input to a lumped hydrological model. This is mostly achieved by considering the average areal rainfall time series, P, equal to the weighted average value of the set of precipitation time series. Therefore, if a weight value w_i is assigned to each station, the following simple equation holds,

$$P = \sum_{i=1}^{N} w_i P_i \tag{1}$$

where P_i is the precipitation time series of the i-th station and N is the total number of rainfall stations. In that sense, each unique combination of stations' weight values is a different spatial distribution of the precipitation data. Therefore, the spatial distribution selected before the hydrological simulation has a significant effect on the performance of the lumped model, since it affects a major forcing time series. This poses the following question: which spatial distribution of precipitation data is the most suitable for a specific watershed?

A typical method used in many cases is the Thiessen method, according to which Thiessen polygons are drawn to divide the basin into sub-areas exclusively corresponding to a single station. Then, the weight value w_i of each station is proportional to the area of its polygon. An alternative simpler distribution occurs when the weights of all stations have the same value. In that case, equation (1) is simplified to the definition of numeric average, and for i=1…N, it holds w_i=1/N.

Finally, the naïve approach is to select one precipitation time series from a single station as the most representative for the watershed.

Each spatially averaging precipitation method over a watershed can be evaluated by calibrating the model to fit the observed runoff. A method that leads to poor calibration, according to specific criteria, is unsuitable, whereas the opposite is valid. Therefore, a suitable method can be any combination of weights that performs adequately when the calibration of the model is performed.

In the present article, a novel method is proposed to determine spatial distribution of precipitation resulting by maximizing the meaningful mutual information contained in the set of spatially averaged precipitation and observed runoff time series. To achieve this, the concept of mutual information is defined according to the Theory of Information, developed by Claude Shannon (Shannon, 1948). A simple genetic algorithm is employed to search for the spatial distribution i.e., the weight vector that maximizes the mutual information between observed runoff and specially averaged precipitation (Findanis & Loukas, 2022). The observed runoff time series is considered accurate. This method is applied at two watersheds by calibrating and validating five lumped monthly models.

2. CONCEPTS OF INFORMATION THEORY

2.1 Shannon's entropy

In 1948, Claude Shannon (Shannon, 1948) defined the entropy H of a discrete random variable X that adheres to the Probability Mass Function (PMF), P(X), as follows:

$$H = -\sum_{i=1}^{N} P(x_i)\log_2 P(x_i) \tag{2}$$

The values of variable X are represented by x_i, where i ranges from 1 to N. The entropy is measured in *Bits* when the base of the logarithm in equation (2) is 2, in *Nats* when the base is e, and in *Harleys* when the base is 10. Equation (2) defines the entropy of a probability mass function P(X), which can be used to determine the uncertainty of the event controlled by the same PMF. A narrower PMF has lower entropy than a wider one, meaning that less uncertain events have less entropy. In the case where an event has N outcomes that are equally probable, its uncertainty H is equal to:

$$H = \log_2 N \tag{3}$$

Equation (3) is derived from equation (2) by assigning P(xi)=1/N for every i. Furthermore, equation (3) implies that when an observer seeks information about the outcome of an event with equal probabilities, they must look for a minimum of H binary questions to obtain the necessary information. In a uniform PMF, where there is no prior knowledge about the most probable outcome, the entropy is at its maximum. Thus

$$0 \le H(X) \le \log_2 N \tag{4}$$

Note that, a time series is a set of ordered observations that contain information. Since the uncertainty of an event that has already occurred is zero, the information gained by observing it must be equivalent to the entropy of the prior probability density function (pdf) of that event. Hence, recorded events do carry information, and this also holds for a time series, whether it comprises observed events or simulated values.

2.2 Joint entropy of a complex source

Suppose there is an information source that generates N events, drawn from a discrete set X according to a Probability Mass Function (PMF) P_x. This source can be denoted as (X, P_X), and its entropy H(X) can be estimated using equation (2). A hydrological model or a watershed can be a case of such a source. In the same manner, a second information source (Y, P_Y) will have entropy equal to H(Y). The system consisting of sources (X, P_X) and (Y, P_Y) forms another information source, with events belonging to the set XY. This complex source is represented by (XY, P_{XY}), where P_{XY} is the joint probability mass function of variables X and Y. The joint entropy of this discrete complex source is given by the following equation:

$$H(X,Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log_2 P(x_i, y_j) \tag{5}$$

and the following inequality always is valid,

$$H(X) + H(Y) \geq H(X,Y) \tag{6}$$

Equation (6) is an equality if X and Y are independent. In cases in which source X and Y are interdependent, the act of observing events produced by source Y can reduce the uncertainty associated with source X. The entropy of source X after such an observation is referred to as the conditional entropy, denoted by H(X|Y). Mathematically, conditional entropy is defined as:

$$H(X \mid Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log_2 \frac{P(x_i, y_j)}{P(x_i)} \tag{7}$$

2.3 Differential entropy

It is possible to apply the concept of entropy to a continuous random variable X that follows a probability density function p(X). The probability of X falling within a particular interval of width dX is equal to p(X)dX. Therefore, equation (2) can be expressed as:

$$H(X) = -\int_{X} p(X)\left[\log_2 p(X)\right] dX - \log_2 dX \tag{8}$$

or

$$H(X) = h(X) - \log_2 dX \tag{9}$$

The integral of equation (8) is referred to as differential or continuous entropy and is denoted by h(X). Unlike Shannon entropy, h(X) does not represent a valid measure of uncertainty and holds no physical meaning. Equation (9) implies that the Shannon entropy of a continuous density distribution is infinite, since $\log_2 dX$ equals to -∞. However, the differential entropy of a continuous variable is finite.

For two continuous variables X and Y, the bivariate joint differential entropy is given by the equation:

$$h(X,Y) = -\int_{X}\int_{Y} p(X,Y)\left[\log_2 p(X,Y)\right] dX dY \tag{10}$$

Thus, for bivariate differential entropy h(X,Y) and the respective Shannon entropy H(X,Y), the following relationship holds:

$$H(X,Y) = h(X,Y) - \log_2 dX - \log_2 dY \tag{11}$$

2.4 Bivariate mutual information

Two variables X and Y, which are statistically dependent, contain mutual information. For instance, time series of precipitation and runoff, contain mutual information because precipitation causes runoff. Mutual Information between variables X and Y is denoted by I(X,Y) and it is a non-negative quantity because of equation (6). In the case that X and Y are discrete variables,

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{12}$$

whereas, if X and Y are continuous variables,

$$I(X,Y) = h(X) + h(Y) - h(X,Y) \tag{13}$$

Furthermore, when X and Y are statistically independent variables, they do not share mutual information, i.e. I(X,Y)=0. Note that mutual information does not vary under reparameterization of the variables (Kraskov et al. 2004). Hence, if $X' = F(X)$ and $Y' = G(Y)$, then $I(X,Y) = I(X',Y')$.

Figure 1 displays a Venn diagram representing a pair of informational sources. The area where the two sources overlap can be considered as the intersection of the two sources, which represents mutual information. On the other hand, the entire area covered by both sources represents their union, which represents joint entropy.

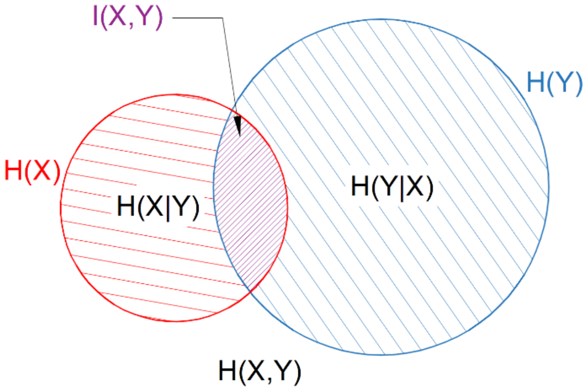

Figure 1. Depiction of sources X and Y. The marginal entropy of X is represented by the red circle, while the blue circle represents the marginal entropy of Y. The mutual information of variables X and Y is indicated by the intersection of the two circles. Similarly, the joined entropy of X and Y is represented by the union of the two circles. The areas that exclusively belong to one variable symbolize the conditional entropies of X given Y H(X|Y) and of Y given X H(Y|X).

2.5 Multivariate mutual Information

Runoff at a particular time step depends on previous time steps of the precipitation time series. Bivariate mutual information considers only the dependence between runoff and precipitation in the same instance of time. For this reason, a more realistic measure of information content can be obtained by computing the multivariate mutual information between lagged time series of rainfall and runoff. A lagged version of a time series has the values of the original time series shifted by a time step. If the precipitation time series is P={P_0, P_1, P_2,… P_N}, we define its lagged version as

P^{t-1}={P_1, P_2, P_3,... P_{N+1}}. The multivariate mutual information between the set of time series **X**={P, P^{t-1}, ... P^{t-C}} and runoff Y, where C is an integer, is defined as,

$$I(\mathbf{X};Y) = h(\mathbf{X}) + h(Y) - h(\mathbf{X},Y) \tag{14}$$

Relationship (14) is similar to equation (13) with the difference that h(**X**) and h(**X**,Y) are multivariate differential entropies because set **X** contains C+1 time series: the integrated precipitation time series P and C lagged versions of it.

To estimate a multivariate differential entropy, an Independent Component Analysis is needed to find the mixing matrix **A** and the set **S** of independent signals (Gong et al., 2013) that satisfy the following linear transformation:

$$\mathbf{X} = \mathbf{S} \cdot \mathbf{A}^T \tag{15}$$

If **S** and **A** are found, then the multivariate differential entropy is estimated as (Cover & Thomas, 2006):

$$h(\mathbf{X}) = h(\mathbf{S}) + \log_2 \left|\det(\mathbf{A})\right| \tag{16}$$

Since signals S are independent, it is:

$$h(\mathbf{S}) = \sum_i h(S_i) \tag{17}$$

Different mixing matrices **A** can satisfy equation (15), leading to slightly different values of h(**X**). Thus, computation of multivariate mutual information can be unstable because ICA must be performed twice, to estimate h(**X**) and h(**X**,Y). To alleviate this problem multiple estimations of I(**X**;Y) are performed and an average value is accepted as its true value.

2.6 Computing univariate and bivariate differential entropies

The main advantage of differential entropy is having a finite value for continuous variables, like rainfall or runoff, whereas Shannon Entropy is infinite for continuous variables. By definition, for the univariate case,

$$h(X) = -\int_X p(X)\left[\log_2 p(X)\right] dX \tag{18}$$

Hence, the right-hand integral must be computed without explicitly knowing the probability density function of X, which is inferred from the observed values of the respective time series. According to Gupta and associates (Gupta et al., 2021), by dividing the domain of X into M equiprobable intervals, whose edges are the quantiles of time series X, the following approximation occurs:

$$p(X)\Delta x_i \approx 1/M \tag{19}$$

where Δx_i is the width of the i-th interval. The integral of equation (18) is approximated as a finite sum. Thus,

$$h(X) \approx -\sum_{i=1}^{M} p(X)\Delta x_i \log_2 p(X) \tag{20}$$

Due to relationship (19), equation (20) may be re-written as:

$$h(X) \approx \log_2 M + \frac{1}{M}\sum_{i=1}^{M} \log_2 \Delta x_i \tag{21}$$

Equation (21) is used to estimate univariate differential entropy by locating the quantiles of X.

Similarly, bivariate differential entropy can be computed by dividing the 2D domain into two-dimensional equiprobable bins using an adaptive grid scheme (Hoshen et al. 2013). If L+1 is the number of quantiles along both axes X and Y, implying that the total number of two-dimensional bins is L^2, the following approximation holds:

$$p(X,Y)\Delta x_i \Delta y_j = 1/L^2 \tag{22}$$

because each portion of the domain has length Δx_i, width Δy_i, and probability of occurrence $1/L^2$. Hence, equation (10) is written as:

$$h(X,Y) \approx \log_2 L^2 + \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} \log_2 \Delta x_i \Delta y_j \tag{23}$$

As a result, the utilization of the adaptive grid method enables the estimation of the bivariate differential entropy h(X,Y) through the application of equation (23).

3. APPLICATION

3.1 Description of hydrological basins

The Pinios River in Thessaly, Greece, originates from the Pindos mountain range and flows into the Aegean Sea. This paper focuses on studying two neighboring sub-watersheds of the Pinios River basin, Mouzaki and Pili watersheds, which have areas of 144.1 km^2 and 132.2 km^2, respectively (Figure 2). These neighboring watersheds have similar geomorphic and hydroclimatic characteristics and their relatively small size makes them suitable for the use of lumped hydrological models. Both basins do not receive runoff from upstream basins. The mean monthly temperature in this area varies significantly throughout the year, ranging from a minimum temperature of -2°C to a maximum temperature of 30°C. The average annual precipitation in the area is about 1400 mm, but it is unevenly distributed spatially and temporally. Mouzaki watershed has an average, maximum, and minimum elevation of 816 m, 1972 m, and 194 m, respectively, whereas, Pili watershed has an average, maximum, and minimum elevation of 957 m, 1872 m, and 264 m, respectively. The mean annual runoff of Mouzaki watershed is about 826 mm, For Pili watershed, the mean annual runoff equals 1128 mm. Both basins are predominantly covered by forests, meadows, and cultivated areas, with urban areas covering an inconsequential percentage of their total area.

3.2 Available data

Monthly runoff time series are available at the outlet of Mouzaki watershed for the period of 10/1960 – 9/1994, except for a missing data period ranging from 10/1985 to 9/1987. Additionally, monthly measurements of precipitation were recorded at five (5) stations for the same period, whereas monthly temperature was measured at the Argithea station. The position and elevation of these stations are presented in Table 1. The Thiessen method was employed to partition the surface of the basin into sub-areas (Thiessen polygons), with each sub-area dominated exclusively by one precipitation station. Table 1 displays the surface of every Thiessen polygon and its corresponding proportion in relation to the overall watershed area. Similarly, monthly observations of runoff are available for the entire period of 10/1960 – 9/1993 at the outlet of Pili watershed. Monthly precipitation and temperature are measured at six and three stations, respectively, for the same period, as shown in Tables 2 and 3.

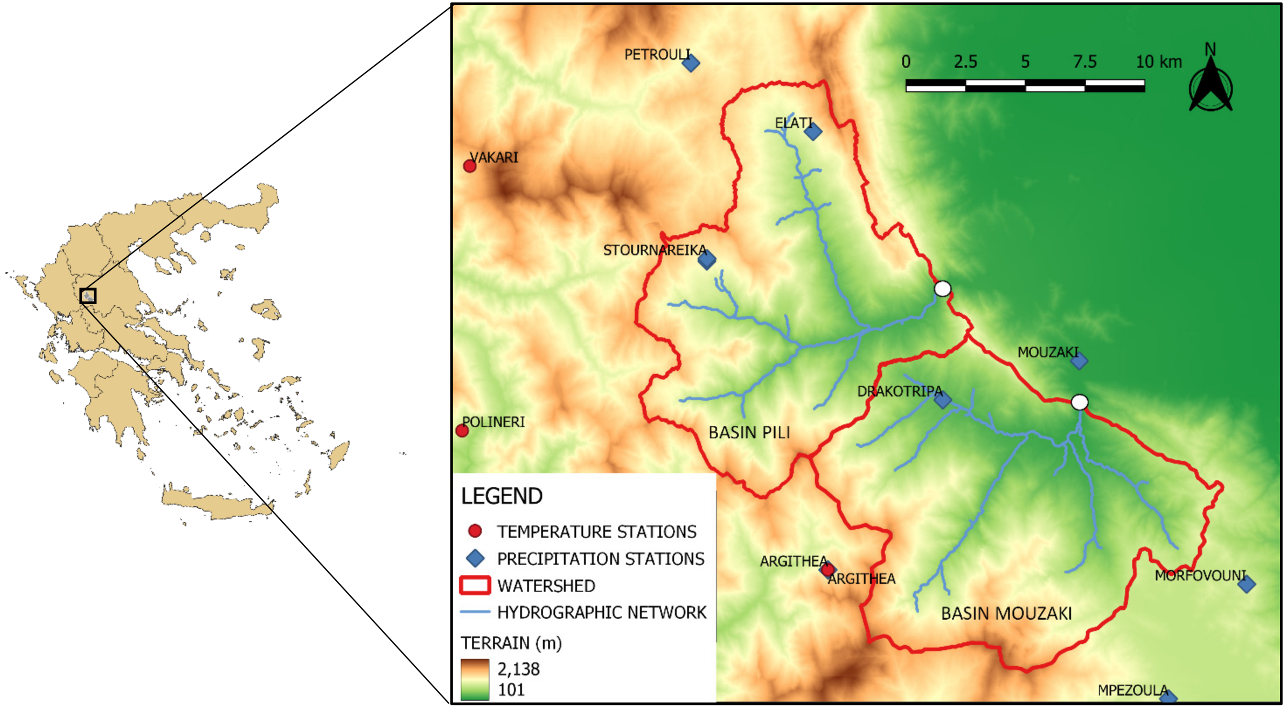

Figure 2. Map of Mouzaki and Pili watersheds and the location of meteorological stations. The outlets of the watersheds are marked by white dots (Findanis & Loukas, 2022).

The observed data, containing records of runoff, precipitation, and temperature, for both watersheds, are divided into two separate data sets. The first data set covering the period of 10/1960-09/1982 is employed for the calibration of the hydrological models. The second data set consisting of the remaining records is utilized to validate the models. Thus, the validation period for Mouzaki watershed ranges from 10/1982 to 09/1994, whereas for Pili watershed ranges from 10/1982 to 09/1993.

Table 1. The rainfall stations inside or near Mouzaki watershed.

Rainfall Station	X (GGRS87)	Y (GGRS87)	Z (m)	Thiessen Polygon Area (km^2)	Percentage of Thiessen area (%)
Argithea	288367.56	4358234.50	980	34.62	24.1
Drakotripa	293185.00	4365363.00	680	52.11	36.3
Mpezoula	302639.31	4352821.00	901	17.45	12.2
Morfovouni	305915.00	4357630.00	780	17.08	11.9
Mouzaki	298900.00	4367000.00	226	22.35	15.6

Table 2. The rainfall stations inside or near Pili watershed.

Rainfall Station	X (GGRS87)	Y (GGRS87)	Z (m)	Thiessen Polygon Area (km^2)	Percentage of Thiessen area (%)
Argithea	288367.56	4358234.50	980	10.63	8.0
Drakotripa	293185.00	4365363.00	680	29.75	22.5
Elati	287748.00	4376618.00	900	30.88	23.4
Petrouli	282626.50	4379493.00	1160	1.21	0.9
Stournareika A	283294.00	4371187.00	860	47.25	35.7
Stournareika B	283300.00	4371287.00	860	12.50	9.5

Table 3. The temperature stations inside or near Mouzaki and Pili watersheds. The Thiessen polygon area and the respective percentage corresponds to Pili watershed.

Temperature Station	X (GGRS87)	Y (GGRS87)	Z (m)	Thiessen Polygon Area (km^2)	Percentage of Thiessen area (%)
Argithea	288367.56	4358234.50	980	76.69	58.5
Vakari	273365.00	4375174.00	1150	45.94	35.0
Polineri	273040.00	4364074.00	730	8.48	6.5

3.3 Evaluation of spatial distribution scenarios

In Tables 4 and 5 the spatially averaged methods (or distribution scenarios) examined for Mouzaki and Pili watersheds, respectively, are presented. For each scenario, the information Ω contained in the integrated dataset is estimated and the bivariate mutual information between the integrated rainfall time series P and the observed runoff Q is calculated. Ω is the multivariate mutual information between P, P^{t-1}, P^{t-2}, P^{t-3}, P^{t-4} and Q, i.e.,

$$\Omega = \mathrm{I}(\mathbf{P};\mathrm{Q}) = \mathrm{h}(\mathbf{P}) + \mathrm{h}(\mathrm{Q}) - \mathrm{h}(\mathbf{P},\mathrm{Q}) \tag{24}$$

where $\mathbf{P}$={P, P^{t-1}, P^{t-2}, P^{t-3}, P^{t-4}} is the set containing integrated time series P and its lagged versions. For the present study, it is assumed that runoff at time step $t=t_0$ does not depend on the rainfall at time steps $t<t_0-4$, i.e. C=4. Additionally, for each basin, equation (14) is used to compute the information A of the dataset consisted of the rainfall time series of all stations and the runoff observed time series at the outlet. It holds,

$$\mathrm{A} = \mathrm{I}(\mathbf{P}_1,...\mathbf{P}_\mathrm{N};\mathrm{Q}) \tag{25}$$

where $\mathbf{P}_i$={P_i, P_i^{t-1}, P_i^{t-2}, P_i^{t-3}, P_i^{t-4}} is the set including the precipitation time series of a specific station and its four lagged versions. For Mouzaki (N=5), it is estimated that A=1.296 bits. For Pili (N=7), A=1.003 bit. Hence, A is the information of the dataset before its compression, Ω is the information of the compressed dataset, the difference A-Ω is the information lost due to the act of integration and the ratio Ω/A is the compression ratio. In every case, Ω/A<1 suggesting that the compression of the original precipitation dataset {$\mathbf{P}_1$,…, $\mathbf{P}_\mathrm{N}$, Q} into the integrated dataset {P,Q} is lossy: the original dataset cannot be recovered if only {P,Q} is known.

Weights w_1 to w_5 in Table 4 represent the effect of the five precipitation stations mentioned in Table 1 on the integrated precipitation for Mouzaki watershed. In the first scenario, these weights are equal to the percentages of the basin according to the Thiessen polygons. Scenario 2 uses the simple arithmetic average method. In scenarios 3 to 7, the areal averaged precipitation is equal by selecting a certain single precipitation station as representative of the spatially averaged precipitation over the watershed area. The weights of scenario 8 occurred by maximizing the bivariate mutual information I(P,Q), which is a function of weights w_1 to w_5 and serves as an approximation of Ω since it can be optimized more easily. Conversely, minimizing I(P,Q) yields the scenario 9, in which the dominant stations, Argithea and Drakotrypa, have a low weight value.

In Table 5, the weights of spatial distribution scenarios for Pili watershed are presented. Weights w_1 to w_6 correspond to the rainfall stations in Table 2, while w_7 represents the effect of the Mouzaki precipitation station from Table 1, included as a control station. Since the Mouzaki station is a distant station from Pili watershed, it is known that it should not affect the runoff. Consequently, the hydrological models are expected to perform poorly in scenario 9. Scenario 1 uses the Thiessen polygon method, and scenario 2 employs the numeric average method. Scenarios 3 to 9 correspond to selecting only one rainfall station. Scenarios 10 and 11 arise from maximizing and minimizing the function I(P,Q), respectively. Note that Ω and I(P,Q) do not necessarily share the same optima, and in scenario 11, the genetic algorithm used for optimizing I(P,Q) failed to find its global minimum, since scenarios 4, 5 and 9 have a lower value of I(P,Q). For Pili watershed, the temperature time series used as input for the models is obtained by applying the Thiessen method to the temperature stations listed in Table 3. No other spatially averaging methods for temperature has been applied.

Table 4. Weights of areal averaging methods and values of the information functions for each scenario for Mouzaki watershed.

Scenario #	w_1	w_2	w_3	w_4	w_5	Ω (bits)	I(P,Q) (bits)	A-Ω (bits)	Ω/A
1	0.241	0.363	0.122	0.119	0.156	0.989	0.526	0.307	0.763
2	0.200	0.200	0.200	0.200	0.200	1.004	0.515	0.292	0.775
3	1	0	0	0	0	1.000	0.342	0.296	0.772
4	0	1	0	0	0	0.644	0.227	0.652	0.497
5	0	0	1	0	0	0.822	0.273	0.474	0.634
6	0	0	0	1	0	0.375	0.030	0.920	0.290
7	0	0	0	0	1	0.515	0.259	0.781	0.397
8	0.385	0.267	0.074	0.048	0.226	1.085	0.646	0.210	0.838
9	0.000	0.007	0.019	0.282	0.692	0.501	0.304	0.794	0.387

Table 5. Weights of areal averaging methods and values of the information functions for each scenario for Pili watershed.

Scenario #	w_1	w_2	w_3	w_4	w_5	w_6	w_7	Ω(bits)	I(P,Q) (bits)	A-Ω (bits)	Ω/A
1	0.080	0.225	0.234	0.009	0.357	0.095	0	0.877	0.567	0.125	0.875
2	0.167	0.167	0.167	0.167	0.167	0.167	0	0.808	0.505	0.195	0.805
3	1	0	0	0	0	0	0	0.815	0.320	0.187	0.813
4	0	1	0	0	0	0	0	0.394	0.192	0.608	0.393
5	0	0	1	0	0	0	0	0.506	0.183	0.497	0.505
6	0	0	0	1	0	0	0	0.520	0.372	0.483	0.519
7	0	0	0	0	1	0	0	0.897	0.355	0.106	0.894
8	0	0	0	0	0	1	0	0.725	0.341	0.278	0.723
9	0	0	0	0	0	0	1	0.282	0.030	0.721	0.281
10	0.083	0.286	0.097	0.098	0.226	0.159	0.051	0.859	0.687	0.144	0.857
11	0.000	0.000	0.000	0.664	0.000	0.001	0.335	0.504	0.213	0.498	0.503

3.4 Calibration and validation of models

Having estimated Ω, for each combination of scenarios and basins, the lumped monthly parametric models, UTHBAL (Loukas et al., 2007), WBM (Xiong and Guo, 1999), Giakoumakis (Giakoumakis et al., 1991), Abulohom (Abulohom et al., 2001), and GR2M (Mouelhi, 2003) have been used in this study, The hydrological models have been calibrated and validated for the two study basins to examine the relationship between Ω and the performance of the models. The necessary inputs for these models are precipitation, temperature, and potential evapotranspiration time series. The Thornthwaite method was used to estimate potential evapotranspiration. UTHBAL and Abulohom models have of five (5) parameters, while WBM, Giakoumakis, and GR2M models have two (2) parameters. Note that a snow accumulation and snowmelt algorithm developed by Loukas and associates (Loukas et al., 2007) and based on the work of Semadeni-Davies (Semadeni-Davies, 1997) was integrated into each model, increasing their parameters' number by one (1).

All models have been calibrated using a simple genetic algorithm. This algorithm searches in the parametric domain for the point that maximizes a selected objective function, by imitating the evolution and the adjustment of species to their natural environment. Simultaneously, the algorithm protects the best solutions to the optimization problem from fading during the passage of generations. The objective function, employed in the present study for calibrating the models, is the Nash-Sutcliffe Efficiency (NSE). After calibrating the models, the NSE corresponding to the validation period will be evaluated. Calibration and validation results are presented in Tables 6 and 7 for Mouzaki and Pili watersheds, respectively.

Table 6. Calibration and validation results of the five hydrological models for Mouzaki watershed.

Scenario #	UTHBAL		WBM		Giakoumakis		Abulohom		GR2M	
	Cal	Val	Cal	Val	Cal	Val	Cal	Val	Cal	Val
1	0.708	0.703	0.694	0.703	0.648	0.530	0.663	0.626	0.723	0.702
2	0.699	0.736	0.677	0.715	0.632	0.547	0.649	0.662	0.736	0.734
3	0.732	0.664	0.726	0.699	0.641	0.537	0.655	0.610	0.733	0.681
4	0.551	0.604	0.499	0.528	0.400	0.265	0.412	0.343	0.488	0.438
5	0.514	0.666	0.479	0.659	0.429	0.400	0.435	0.525	0.548	0.600
6	0.508	0.389	0.372	0.618	0.258	0.402	0.303	0.499	0.552	0.494
7	0.642	0.612	0.585	0.617	0.452	0.292	0.491	0.457	0.617	0.560
8	0.736	0.618	0.727	0.719	0.691	0.536	0.707	0.655	0.748	0.722
9	0.602	0.674	0.562	0.647	0.459	0.388	0.490	0.493	0.648	0.597

Table 7. Calibration and validation results of the five hydrological models for Pili watershed.

Scenario #	UTHBAL		WBM		Giakoumakis		Abulohom		GR2M	
	Cal	Val	Cal	Val	Cal	Val	Cal	Val	Cal	Val
1	0.725	0.720	0.702	0.727	0.673	0.621	0.700	0.703	0.731	0.758
2	0.704	0.724	0.678	0.746	0.665	0.631	0.673	0.708	0.728	0.778
3	0.683	0.698	0.676	0.707	0.617	0.590	0.656	0.623	0.694	0.738
4	0.514	0.594	0.475	0.571	0.374	0.292	0.401	0.378	0.471	0.527
5	0.693	0.770	0.675	0.749	0.581	0.581	0.627	0.652	0.679	0.765
6	0.534	0.662	0.497	0.651	0.439	0.446	0.444	0.566	0.581	0.684
7	0.664	0.574	0.662	0.613	0.574	0.437	0.620	0.493	0.662	0.597
8	0.545	0.534	0.520	0.487	0.491	0.295	0.497	0.283	0.570	0.257
9	0.541	0.598	0.447	0.553	0.337	0.323	0.360	0.369	0.561	0.651
10	0.687	0.714	0.662	0.722	0.643	0.606	0.651	0.677	0.713	0.755
11	0.544	0.613	0.499	0.638	0.433	0.471	0.448	0.547	0.614	0.711

4. DISCUSSION

The results in Tables 6 and 7 indicate that the Thiessen polygon method (Scenario 1) leads to high values of NSE for the calibration and validation period of almost all models for both study watersheds. Scenario 2, the mean arithmetic method, may outperform the Thiessen method, especially during the validation period. For Mouzaki watershed, scenario 8 corresponding to the maximum value of Ω, has the highest values of NSE for the calibration period of each model. On the contrary, this is not observed for scenario 7 of Pili watershed, where only the Stournareika A precipitation station is considered and Ω is maximum, because all models perform adequately. This may occur since, at a fundamental level, Ω and NSE express, respectively, the input and the output of models. Hence, their relationship is not linear, but it depends on the structure of models and the value of their parameters.

Moreover, selecting only one precipitation station does not guarantee a good fit of simulated to observed runoff. To be more specific, models perform satisfactorily for the single-station scenario 3 for both study watersheds, where only the Arghithea precipitation station is considered, and for scenario 5 of basin Pili, where the average areal precipitation is equal to the precipitation of Elati station, because, in these two scenarios Ω, is relatively high. But in other single-station scenarios, like scenarios 4, 5, 6 of Mouzaki and 4, 9 of Pili watershed, models perform poorly, due to the low value of meaningful information Ω. In scenario 8 of Pili watershed, all models exhibit their worst performance, compared to the rest scenarios of the same watershed, while their inputted information is not too low (Ω=0.725 bits). This happens for the same reason why models do not show their best performance for scenario 7 of Pili: Although Ω drives NSE, the model structure and its selected parameters define their exact relationship. Figure 3 presents the results of Tables 6 and 7.

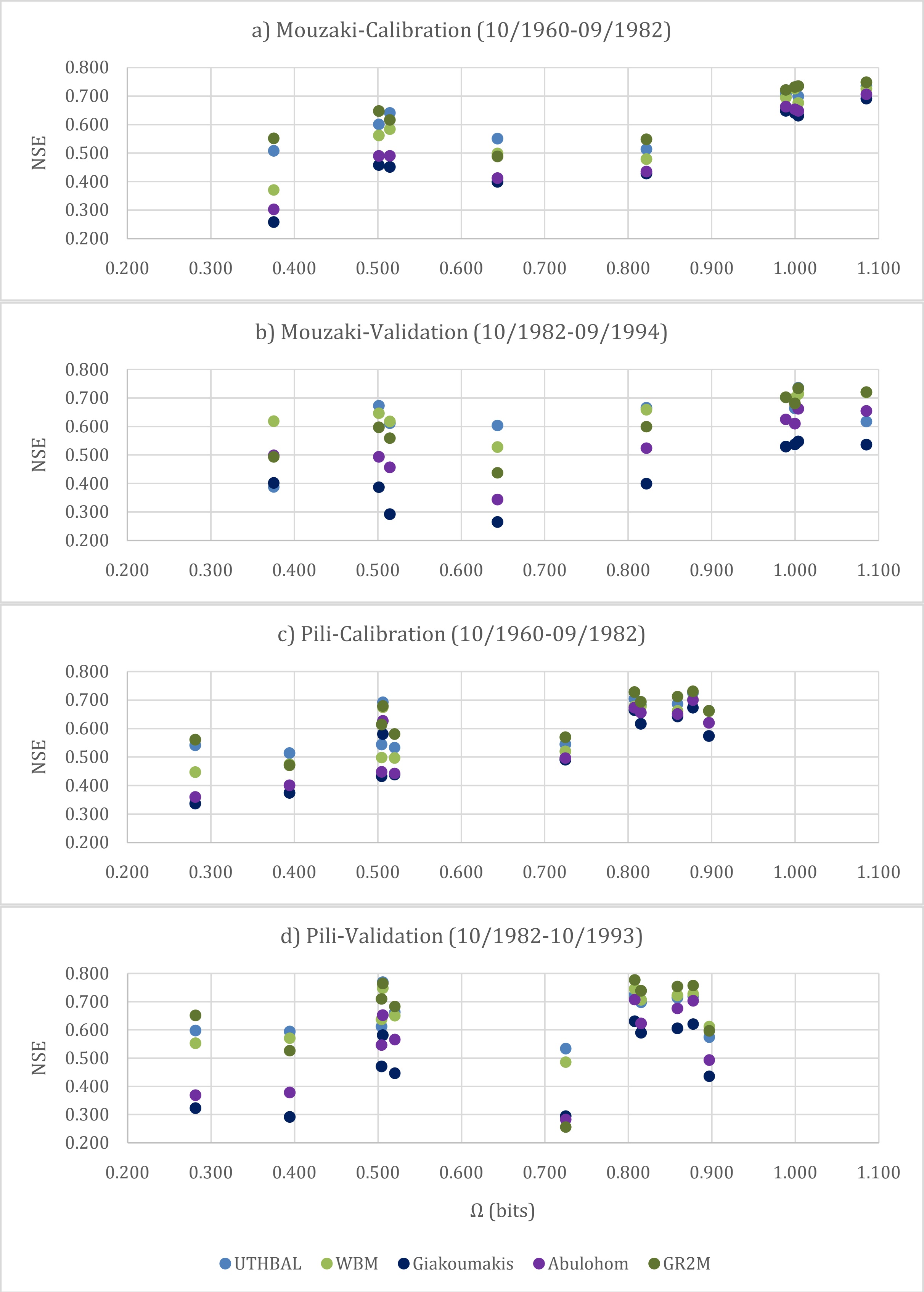

Figure 3. NSE criterion versus information Ω for calibration and validation periods of Mouzaki and Pili watersheds.

To summarize, the relationship Ω~NSE exists but it is weak and non-linear. This is true because: a) NSE describes models' performance, whereas Ω measures their information input, b) An accurate calculation of Ω is challenging due to the required Independent Component Analysis, c)

Information due to temperature variation was ignored. The main conclusion from Figure 3 is that a high value of Ω leads to satisfactory NSE values, but a low value of Ω does not necessarily correspond to low NSE values. Thus, a spatially averaging method of precipitation data, which outperforms the Thiessen polygon method, can occur by maximizing Ω.

The main advantage of the Thiessen polygon method over maximizing Ω is its simplicity. Maximizing Ω is challenging, requires an optimization algorithm, and cannot be done directly because the algorithm optimizes the function I(P,Q), not Ω itself. On the other hand, using Ω to construct spatial distribution scenarios of precipitation has more physical meaning: The hydrologist tries to construct a single time series of precipitation which is the most valuable in terms of information. Also, the proposed methodology may give an alternative, more theoretical, interpretation of empirical facts. For instance, in scenario 6 of Mouzaki watershed and scenario 9 of Pili watershed, where Ω is minimum and distant stations with small or zero Thiessen percentages have weight values equal to one, all models perform poorly as expected. This can be interpreted empirically as "distant precipitation stations should not affect the runoff" or accordingly to the proposed framework as "distant stations deprive models of useful information".

Finally, models UTHBAL, WBM, and GR2M show high NSE values for both calibration and validation periods if enough information is provided to them. Giakoumakis is the model with the worst performance from all hydrological models used.

ACKNOWLEDGMENTS

An initial shorter version of the paper has been presented in the 15th Conference of the HHA, Thessaloniki, Greece, 2-3 June 2022.

This work was supported by the Academy of Athens.

REFERENCES

Abulohom M.S., Shah S.M.S. and Ghumman A.R., Development of a Rainfall-Runoff Model, its Calibration and Validation, Water Resources Management 15(3), 149–163, 2001, https://doi.org/10.1023/A:1013069709740

Cover T. and Thomas J., Entropy, Chapter 2: Relative Entropy and Mutual Information, Elements of Information Theory, 2nd Edition, 2006, ISBN: 978-0-471-24195-9

Findanis E., Loukas A., The effect of spatial distribution of precipitation data on hydrological simulation with monthly lumped models, 15th Conference of the HHA, Thessaloniki, Greece, 2 – 3 June 2022

Giakoumakis S., Tsakiris G., and Efremides D., On the Rainfall-runoff Modeling in a Mediterranean Island Environment, Advances in Water Resources Technology, Balkema, Rotterdam, 1991

Gong, W., H. V. Gupta, D. Yang, K. Sricharan, and A. O. Hero III, Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach, Water Resources Research, 49, 2253–2273, 2013, https://doi.org/10.1002/wrcr.20161.

Gupta H.V., Ehsani M.R., Roy T., Sans-Fuentes M.A., Ehret U. and Behrangi A, Computing Accurate Probabilistic Estimates of One-D Entropy from Equiprobable Random Samples. Entropy 23(6), 740, 2021, https://doi.org/10.3390/e23060740

Hoshen Y., Arora C., Poleg Y. and Peleg S., Efficient representation of distributions for background subtraction, 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, 276-281, 2013, doi: 10.1109/AVSS.2013.6636652

Kraskov A., Stögbauer H., and Grassberger P., Estimating mutual information. Physical Review E, 69(6), 066138, 2004, https://doi.org/10.1103/PhysRevE.69.066138

Lihua Xiong and Shenglian Guo, A two-parameter monthly water balance model and its application, Journal of Hydrology 216(1–2), 111-123, 1999, https://doi.org/10.1016/S0022-1694(98)00297-2

Loukas A., Mylopoulos N. and Vasiliades L, A Modeling System for the Evaluation of Water Resources Management Strategies in Thessaly, Greece, Water Resources Management 21, 1673–1702, 2007, https://doi.org/10.1007/s11269-006-9120-5

Mouelhi S., Vers une chaîne cohérente de modèles pluie-débit conceptuels globaux aux pas de temps pluriannuel, annuel, mensuel et journalier. PhD thesis (in French), ENGREF, Cemagref Antony, France, 2003

Semadeni-Davies A., Monthly snowmelt modeling for long-scale climate change studies using the degree-day approach, Ecological Modelling, 101(2-3), 303-323, 1997, https://doi.org/10.1016/S0304-3800(97)00054-9

Shannon C.E., A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379-423, 1948, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x