Environmental Modeling & Assessment (2022) 27:817–830
https://doi.org/10.1007/s10666-022-09838-1

# Conceptual Stormwater Quality Models by Alternative Linear and Non-linear Formulations: an Event-Based Approach

**Santiago Sandoval**[1,2,3] · **Jean-Luc Bertrand-Krajewski**[1] · **Felipe Peña-Heredia**[3]

Received: 5 August 2020 / Accepted: 27 May 2022 / Published online: 30 June 2022


**Abstract**
Different linear transfer functions (TFs) models are compared to the traditional non-linear rating curve (RC) model in a novel event-based exploratory frame. This approach aims to scrutinize for alternative representations of the stormwater TSS load dynamics at the output of a 185-ha urban catchment, when not considering explicitly a virtual state of available pollutant mass accumulated during the dry weather and subsequently washed by rainfall (as a simplification of the widespread accumulation/wash-off model). The advantages of using TFs remain unproved compared to the RC model, which shows more performant results in terms of parsimony and simulation uncertainties with a dataset of 255 calibration and 110 verification online-monitored rainfall events. This can be explained by the non-linear properties of RC. Modeling performance when rainfall is used as the input remains lower than for the case of flow rate (from mean NS of 0.6 to 0.4 in verification events). On the other hand, the RC by itself can be still considered an unsatisfactory model (e.g., NS about 0.6 in verification with the flow rate as input), suggesting the lack of an essential missing process in this formulation. However, this missing process (if there is one) might not be necessarily linked to a time-decreasing state of available pollutant mass.

**Keywords** Accumulation/wash-off model · Bayesian calibration · Model selection · Online monitoring · Transfer functions · Water quality modeling

## 1 Introduction

During the last five decades, modeling the dynamics of stormwater total suspended solids (TSS) loads at the outlet of urban catchments has been mainly addressed by the concept of accumulation/wash-off [1]. These model structures are based on a state variable of total available pollutant mass $M$, accumulated during dry weather (build-up) and subsequently washed and transported by stormwater flow rate $Q$ to the outlet of the system. However, accumulation/wash-off models frequently exhibit unsatisfactory performances when implemented in real-world applications, especially with large and complex urban catchments (e.g., [2]). This can be explained by the non-generalizable nature of the accumulation/wash-off concept to larger scales, where a virtual state variable of available pollutant mass $M$ representative for the entire catchment might be inexistent [3, 4]. Alternatively, researchers have also argued about the benefits of adopting a simplified rating curve (RC) model [5], without including explicitly the pollutant mass $M$ process in the calculations (e.g., [6]). Unexpectedly, studies testing further "inexistent $M$" models apart from the RC remain limited.

On the other hand, there has been an increasing interest in linear transfer function (TF) models into environmental sciences due to their physical interpretability [7]. However, implementations in the context of water quality are not numerous (e.g., [8, 9]). Therefore, the TFs emerge as a promising research direction towards alternative descriptions, in the sense that all of them are monotonic functions of the input dynamic, and no virtual state variables are introduced in the calculation (as for the original RC). One immediate difference between TFs and RC is that the second has a well-known physical meaning, linking shear stress produced by flow rate $Q$ to TSS production (due to resuspension) [10]. Moreover, the TFs models are in principle

✉ Felipe Peña-Heredia
felipe.pena@javeriana.edu.co

[1] Université de Lyon, INSA Lyon, DEEP, Lyon, France

[2] University of Applied Sciences and Arts of Western Switzerland (HES-SO), HEIA-Fr, ITEC, Boulevard de Pérolles 80, Fribourg 1700, Switzerland

[3] Pontificia Universidad Javeriana, Bogotá, Colombia

black-box descriptions. However, system identification and control literature enhance their potential to be interpreted as serial, parallel, and feedback connections of sub-systems that often have a physical meaning [8].

Therefore, the objective of this work is to explore different "inexistent *M*" models (TFs and RC) aimed to simulate TSS load pollutographs at the outlet of a separate sewer in a 185-ha urban catchment (Chassieu, France), employing a 2-min time-step dataset of flow rate and TSS load time series, measured from 2004 to 2011. The potential benefits of the TFs compared to RC are assessed by two novel model selection methodologies, considering the inter-event variability for different performance indicators. Results lead to revisiting some postulates of the traditional accumulation/wash-off concept.

## 2 Materials and Methods

### 2.1 Experimental Setup and Dataset

Three hundred sixty-five rainfall events from the Chassieu urban catchment (France) measured between 2004 and 2011, being part of one of the experimental sites of the OTHU project (Field Observatory for Urban Hydrology –www.othu.org), were used to test the proposed methods. The basin has an industrial area of 185 ha that is drained by a separate storm sewer system, with imperviousness and runoff coefficients of about 0.72 and 0.43, respectively. At the outlet of the catchment, flow rates $Q$ [L/s] and TSS concentrations [mg/L] were measured with a 2-min time-step resolution in a 1.6-m circular concrete pipe. Flow rate $Q$ was estimated from water depth measurements, where relative standard uncertainties of $Q$ values are found to be ranging from 15 to 25%, following the law of propagation of uncertainties (LPU) [12]. Regarding TSS concentrations, turbidity measurements [NTU] from drainage water that is pumped into an off-line monitoring flume in a shelter were converted into TSS concentration values by local calibration functions (see [11]). The estimated TSS concentrations present standard uncertainties ranging from 11 to 30%. For the TSS loads [kg/s] (TSS load = TSS concentration $Q$), relative standard uncertainties are reported to be ranging from 15 to 50%. For higher values of the measurands ($Q$, TSS concentrations, load), higher values of standard uncertainties are presented. Figure 1 shows an example of a time series of $Q$ and TSS concentrations from a rainfall event. The proposed methodology is tested by splitting the measured rainfall events into two datasets for calibration (first 255 events) and verification (last 110 events).

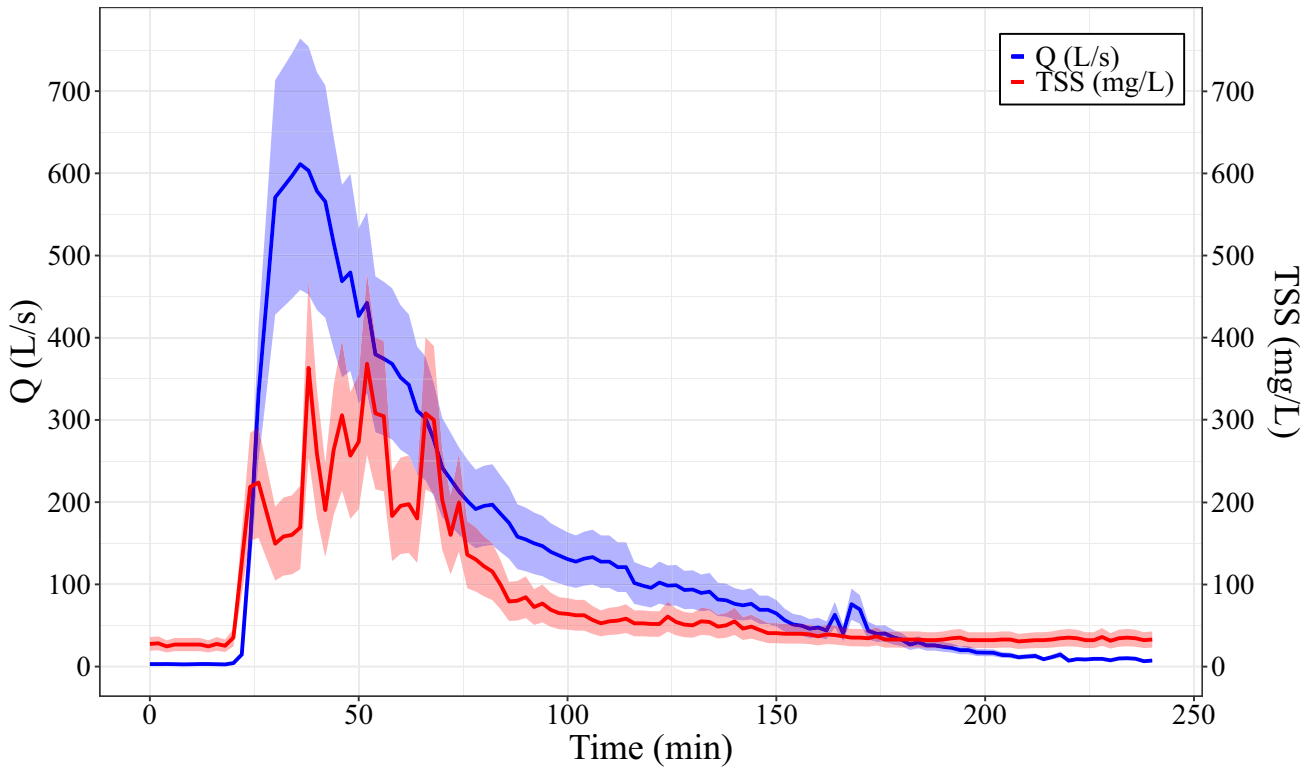


**Fig. 1** Example of a time series of $Q$ (in blue) and TSS concentration (in red) from a rainfall event

### 2.2 Model Structures: Transfer Functions and Rating Curve

A single-input single-output (SISO) TF can be expressed by the following equation:

$$\begin{cases} x(t) = \frac{B(s)}{A(s)} u(t-\tau) \\ load(t) = x(t) + e(t), \end{cases} \quad (1)$$

where

$$A(s) = s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1} s + \alpha_n, \quad (2)$$

$$B(s) = \beta_0 s^m + \beta_1 s^{m-1} + \cdots + \beta_{m-1} s + \beta_m \quad (3)$$

$A(s)$ and $B(s)$ are polynomials with the derivative operator $s = d/dt$ and $\tau$ is a pure time delay, in minutes. In Eqs. (1)–(3), $u(t)$ is the input signal, $x(t)$ is the noise-free output signal, and $load(t)$ is the noisy output signal. The $e(t)$ is considered independent and identically distributed (i.i.d.). However, this assumption is not restrictive for TF applications [13]. A TF is a compact representation of a differential equation. Therefore, the physical interpretability of Eq. (1) can be directly related to the following differential equation from Eq. (4) [8]:

$$\begin{aligned} &\frac{d^n load(t)}{dt^n} + \alpha_1 \frac{d^{n-1} load(t)}{dt^{n-1}} + \cdots + \alpha_n load(t) \\ &= \beta_0 \frac{d^m u(t-\tau)}{dt^m} + \cdots + \beta_m u(t-\tau) + A(s) \cdot e(t). \end{aligned} \quad (4)$$

A given model structure is defined for this type of models from the pair ($n$ poles, $m$ zeros), denoted as TF($n$, $m$). The number of parameters for the set $\theta$ is equal to $n + (m + 1)$, where the set $\theta = [\alpha_1, \ldots, \alpha_n, \beta_0, \ldots, \beta_m]$. On the other hand, the RC [5] remains probably the simplest model in this context, at least compared to the original accumulation/wash-off formulation [1] and further alternatives (e.g., [14, 15]). The RC model is described as follows in Eq. (5):

$$load(t) = M_0 \cdot u(t-\tau)^r, \quad (5)$$

where the set of parameters $\theta = [M_0, r]$. Accordingly, 7 different model structures are tested in this study, including TF(0,0), TF(1,1), TF(2,1), TF(2,2), TF(3,2), TF(3,3), and the RC. For $Q$ as the input $u(t)$, the delay $\tau = 0$ and the number of parameters is: 1, 3, 4, 5, 6, 7, and 2, respectively. For rainfall $R$ as the input $u(t)$, $\tau$ is added as another parameter for all model structures. One difference between TFs a nd RC is their linear and non-linear nature, regardless of the number of parameters. On the other hand, TFs led to include information from previous time-steps (measured u and/or simulated TSS load) in the calculation, contrary to the RC in which the calculation of $load(t)$ only depends on $u(t-\tau)$. Implications of these differences are further discussed in Sect. 3.

### 2.3 Parameter Identification

Although there is a wide range of calibration methods in the literature for TFs, most of them are based on hypotheses related to normality and data input free-error [16]. However, the identification of the set of parameters for each TF or RC model is preferred to be undertaken by an event-based Bayesian approach [17], as proposed in Sandoval and Bertrand-Krajewski [18]. For a given calibration rainfall event $i$, $\theta_i$ is the local set of parameters of a model structure (TFs or RC) and $p(\theta|load)_i$ their probability density function (PDF), given an input $u(t)_{obs_i}$ ($Q$ or $R$) and an output $load(t)_i$. The Bayesian approach leads to calculate local-event-based $p(\theta|load)_i$, named posterior distributions, on the basis of a likelihood function and a prior knowledge of the distribution of parameters $p(\theta)$, which is expressed by Eqs. (6)–(7):

$$p(\theta|load)_i = C_i \prod_{t=1}^{dur_i} \frac{1}{\sqrt{2\pi\hat{\sigma}_t^2}} \exp\left\{ -\frac{1}{2} \frac{\left[res(t)_i\right]^2}{\hat{\sigma}_t^2} \right\} \cdot p(\theta), \quad (6)$$

$$res(t)_i = load_{sim}(t,\theta)_i - load_{obs}(t)_i, \quad (7)$$

where $dur_i$ is the length of the load output data, $load_{sim}(t,\theta)_i$ is the simulated load by a model (TFs or RC) at a given time-step $t$ with a set of parameters $\theta_i$, $p(\theta)$ is a uniform probability distribution for each parameter (informative-less). Prior uniform distributions are aimed to reflect the absence of prior information about the actual values of model parameters. $C_i$ is a normalization coefficient (irrelevant for the implementation of numerical algorithms to solve Eq. (6), see [17]) to guarantee that $\int p(\theta|load)_i d\theta = 1$, $res(t)_i$ are the residuals of the model in Eq. (7) and $\hat{\sigma}_t^2$ is the residual variance, considered for this application to be equal to the squared value of the standard uncertainty of $load_{sim}(t,\theta)_i$ values for each time-step $t$. The DREAM algorithm is used for determining $p(\theta|load)_i$ as a solution to Eq. (6) [17]. The point estimate for the local set of parameters for each event $i$ was adopted as the maximum a posteriori (MAP) of $\theta$ (i.e., the $\theta$ value that reported the maximum likelihood among all samples of $\theta$ from $p(\theta|load)_i$) [17, 19]. This estimate is called $\theta_{map_i}$. The global estimation $p(\theta|load)$ is calculated by the marginalization of representative local estimations of $p(\theta|load)_i$ (events reporting $\theta_{map_i}$ estimations NS < 0.8 are discarded as unrepresentative). The global point estimate of $\theta$ is calculated as the expectation of $p(\theta|load)$, i.e., $\theta_{map} = \int p(\theta|load) \cdot \theta \, d\theta$. It is worth clarification that the adopted Bayesian error model implies that $e(t)$ should be i.i.d. normally distributed, as stated in previous lines. This error model is adopted for the case study in regards of simplicity, as calibrations including this assumption are commonly

**Table 1** Description of indicators *I* explored in this work

| Indicator *I* | Range | Dataset | Source |
|---|---|---|---|
| AIC | From -inf to inf | Calibration | [28] |
| BIC | From -inf to inf | Calibration | [29] |
| YIC*mod* | From -inf to inf | Calibration | [30] |
| NS | From -inf to 1 | Verification | [31] |
| ARIL | From 0 to inf | Verification | [32] |
| POC*mod* | From 0 to 1 | Verification | [33] |

performed in the urban drainage modeling context (e.g., [20, 21]). However, the proposed methodology is fully applicable to cases where likelihood functions are required to include different residuals distributions (e.g., [22]).

## 2.4 Model Selection

The model selection problem is commonly addressed by comparing the scores from a given performance indicator *I* obtained by different candidate models. In the urban drainage context, these indicators *I* are frequently aimed to evaluate models' performance from verification data, where the analyzed models are calibrated beforehand (e.g., [23, 24]). However, this verification stage is not always feasible, due to an insufficient amount of data or high computational requirements. Therefore, alternative statistics *I* have been developed for assessing model appropriateness exclusively from calibration data (e.g., [25, 26]). On the other hand, the selection of the indicators *I* to be analyzed is dependent on the specific objectives of a particular study, whether analyses are carried out with calibration or verification data. For this case study, calibration as well as verification indicators *I* are included with illustrative purposes, which were selected on the basis of recommendations from previous studies and similar settings (e.g., [25, 27]). Specifically, the following indicators *I* are explored: Akaike [28], Bayesian [29], and a modified Young [30] Information Criterion (AIC, BIC, and YIC*mod*, respectively) for calibration data and the Nash–Sutcliffe efficiency (NS) [31], average relative interval length (ARIL) [32], and a modified percentage of coverage (POC*mod*) [33] for verification data (Table 1). This distinction is principally made as NS, ARIL, and POC*mod* are prone to report overestimated performances when calculated with calibration data (overfitting) and AIC, BIC, and YIC*mod* in verification tend to over-penalize the prediction capacity of models with a higher number of parameters [34].

The first step consists in selecting an input $u(t)_{obs}$ between $Q$ or $R$ (Fig. 2a). This selection can be based on the available data and the specific modeling objectives (e.g., [6, 35]). For the implementation as presented here, calculations were carried out for both types of inputs in regards of illustrating the methodology. As mentioned in Sect. 2.2, the candidate model structures $j = 1{:}num\_str$ selected within this implementation are the traditional RC (2 parameters) and the following TFs ($n$ poles, $m$ zeros): (0,0), (1,1), (2,1), (2,2), (3,2), (3,3), for a total of model structures $num\_str = 7$ (Fig. 2b). From Sect. 2.3., local $p(\theta|load)_i$ and $\theta_{\mathrm{map_i}}$ event-based estimations are provided for each of the calibration rainfall events (i.e., $i = 1{:}255$ for this implementation) (Fig. 2c).

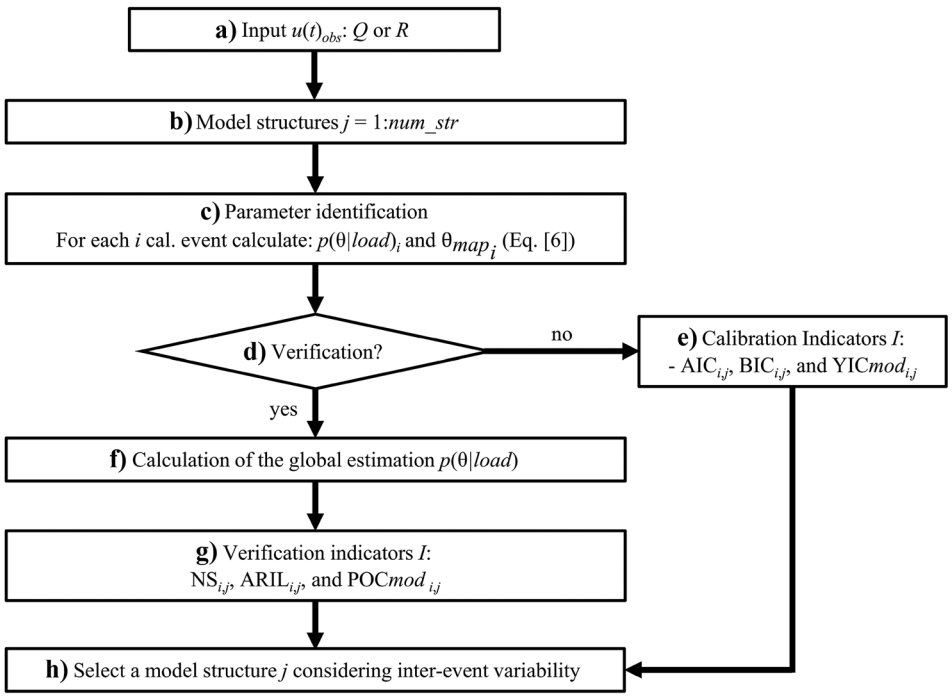


**Fig. 2** Proposed methodology for model selection

In the case that verification is not performed (Fig. 2d), the AIC, BIC, and YIC*mod*, can be evaluated employing the calibration dataset. The Young Information Criterion (YIC) as proposed by Young [30] is a function of the standard deviation $\sigma_\theta$ of the parameters of the set $\theta$. However, the Bayesian approach proposed in Eq. (6) allows to deliver a detailed description of $p(\theta|load)$, which might be neither normal nor symmetric. Hence, parametric uncertainties cannot be summarized by $\sigma_\theta$, limiting the direct application of YIC in its classical form. Therefore, $\sigma_\theta$ in the original YIC indicator is replaced by a non-parametric estimation of dispersion for $p(\theta|load)$. Accordingly, the YIC*mod* is calculated as follows in Eq. (8):

$$\mathrm{YIC}mod = \ln\left[\frac{\sum_{t=1}^{dur_i} res(t)^2}{\sigma^2_{load}}\right] - \ln\left\{\frac{1}{n_p}\sum_{k=1}^{n_p}\frac{p(\theta(k)/load)_{97.5} - p(\theta(k)/load)_{2.5}}{\theta(k)_{map}}\right\}, \tag{8}$$

where $p(\theta(k)|load)_{97.5}$ and $p(\theta(k)|load)_{2.5}$ are the 97.5% and 2.5% limits of the $p(\theta(k)|load)$ distribution, for a 95% coverage of the parameter $\theta(k)$ from the set $\theta$. The $\sigma^2_{laod}$ remains the variance of the load pollutograph, and $n_p$ is the number of parameters of the studied model structure. The first term in the calculation of YIC*mod* in Eq. (8) is simply a relative, logarithmic measure of how well the model explains the data: the smaller the model residuals the more negative the term becomes. The second term of YIC*mod* in Eq. (8) provides a logarithmic measure of the relative uncertainty in the model parameters. If the model is over-parameterized, then it can be shown that the trace of the covariance matrix of the parameters will increase, often by several orders of magnitude (e.g., [13]). When this is the case, the second term in the YIC*mod* tends to dominate the criterion function, indicating over-parameterization. For the case of AIC or BIC, a similar tradeoff is provided between adjustment (first term) and complexity (second term), where complexity/over-parameterization are principally a function of $n_p$, without accounting explicitly for parametric uncertainties [34].

However, neither the AIC, BIC, nor YIC*mod* are aimed to bring direct information about the adjustment of the model itself, as these indicators are exclusively designed for comparative purposes within a set of model structures. Hence, analyses from calibration data by mentioned indicators *I* must be complemented with further information about the adjustment of the model structures. The NS coefficient from calibration is used for this purpose in the present study, stressing that NS should not be employed standalone for calibration, as argued in previous lines.

On the other hand, when performing event-based evaluations, indicator *I* for a model structure *j* will report different values for each rainfall event *i* named $I_{i,j}$, due to the inter-event variability. Accordingly, the $\mathrm{AIC}_{i,j}$, $\mathrm{BIC}_{i,j}$, $\mathrm{YIC}mod_{i,j}$, and $\mathrm{NS}_{i,j}$ are calculated from each local $p(\theta|load)_i$ and $\theta_{\mathrm{map_i}}$ event-based estimation (Fig. 2e). Event-based comparisons among all $I_{i,:}$ (: referring to all candidate models) can lead to provide evidence about a "best" model structure for each calibration rainfall event *i*. For that case, the procedure indicated in Fig. 2h can be directly applied to select a model structure accounting for the inter-event variability, despite the inherent limitations of not implementing a verification stage (e.g., [36]). The model selection procedure from Fig. 2h is described at the end of this Sect. 2.4.

On the other hand, the performance of the models can be further assessed in a verification stage. For this specific implementation, the verification stage is carried out over the remaining $i = 256{:}365$ events, by using the global estimation $p(\theta|load)$ and the global point estimate $\theta_{map}$ previously obtained from the parameters identification from Sect. 2.3. (Fig. 2f). The following indicators $I_{i,j}$ are then calculated for each verification event *i* and a given *j* model structure: the size of uncertainty intervals for load predictions (precision) ($\mathrm{ARIL}_{i,j}$ [37]), the number of measurements inside the uncertainty intervals (reliability) ($\mathrm{POC}mod_{i,j}$ [18]) and the adjustment (accuracy) ($\mathrm{NS}_{i,j}$) (Fig. 2g).

Delivering a conclusion about a "best" model structure *j* (Fig. 2h), based on the corresponding $I_{i,j}$ estimations obtained from events either from calibration or verification, represents additional challenges. This selection is not direct given that the "best" model structure *j* might be different from one event to another. The selection in Fig. 2h is proposed to be carried out by comparing the indicators $I_{:,j}$ (: referring to all calibration or verification events) under two different approaches:

- Approach of selection 1 (S1): the model selection is undertaken based on the idea of evaluating if the inter-event mean (or median) of the set of values reported for an indicator $I_{:,j}$, by a model structure *j*, is significantly higher (or lower) than the inter-event mean (or median) of the *I* scores obtained by any of the other model structures $I_{:,\neq j}$. For this purpose, an analysis of variance (ANOVA) test is proposed as a first step to evaluate the null hypothesis that there is at least one model structure that reports an *I* mean score significantly different from the other models or groups of models, versus the alternative hypothesis that none of the model structures does, for a significance level of 5% ($p$ value $< 0.05$). For the case in which normality cannot be verified for at least one of the $I_{:,j}$ groups (Shapiro–Wilk test, significance level of 5%), the non-parametric Kruskal–Wallis test is implemented as an alternative to ANOVA [38]. For this case, the Kruskal–Wallis test aims to verify the null and alternative hypotheses over the median *I* scores, instead of the means, as for the case of ANOVA. Afterwards, a post hoc Pairwise Tukey's Honest Significant Difference (Tukey's HSD) is applied to identify which is the group $I_{:,j}$ that presents a significantly different mean (or