The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1757–1775
DOI: 10.1214/16-AOAS958

CORRECTION OF BIFURCATED RIVER FLOW MEASUREMENTS FROM HISTORICAL DATA: PAVING THE WAY FOR THE TEESTA WATER SHARING TREATY

BY KAUSHIK JANA*, DEBASIS SENGUPTA* AND KALYAN RUDRA†

*Indian Statistical Institute** *and West Bengal Pollution Control Board*†

In this paper, we consider an estimation problem arising in the measurement of bifurcated flow of the Teesta, a trans-boundary river flowing through India and Bangladesh. The location of measurement is an Indian Barrage, where a part of the flow is diverted from the main stream to a canal. The flows through the two channels are regulated by different control structures and are measured indirectly from the height of the water level and the dimensions of the control structures. The computational formula for the measurement involves a hydrological constant used as a multiplier. Empirical findings indicate that incorrect multipliers are currently used in the computational formula for the two channels. For implementing any water sharing treaty between the two countries, the measurements need to be brought to a common scale. For this purpose, we present a model with carefully considered assumptions to estimate the correction factor. The model permits diagnostic tests for validation of the assumptions. We provide a nonparametric and consistent estimator of the desired factor.

Analysis of historical flow data shows that a main stream flow measured as 100 cumec would be measured as 76 cumec if it is diverted through the canal. Adjustment of emerging measurements through this finding would help the governments of India and Bangladesh to effectively implement and monitor any water sharing agreement.

1. Introduction. Sharing of water of a trans-boundary river has been an issue of concern for neighbouring states for many centuries [Vidal (2010), Solomon (2010)]. A major dispute of current interest concerns the river Teesta, a major trans-boundary river shared by India and Bangladesh. The river originates from the eastern Himalayas in the Indian state of Sikkim, flows through the Indian state of West Bengal and crosses the international border to enter Bangladesh, before merging with the river Brahmaputra in Bangladesh. The Teesta, which supports the ecology of its vast basin and provides key support to agriculture, is regarded as the lifeline of a number of districts of India and Bangladesh [Rudra (2012)]. Sharing of the flow of the Teesta, particularly during the lean season, has been a matter of major and long-standing contention between India and Bangladesh. In the recent past, there have been some indications that an accord on water sharing may

Received December 2015; revised June 2016.

Key words and phrases. Bifurcation, dependence measure, hydrological constant, independence, multiplicative distortion, trans-boundary river.

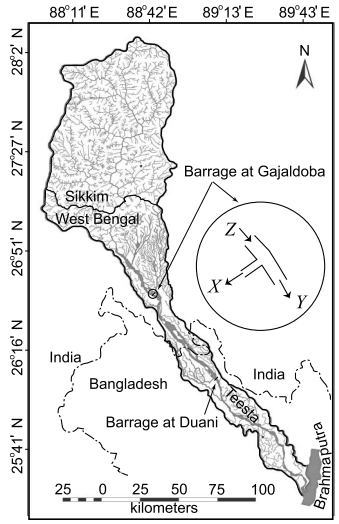

FIG. 1. *Map of the Teesta basin with locations of barrages and the international border* [*source*: *Rudra* (2012) *with permission*].

materialize soon [India Today (2015)]. The anticipated Teesta accord is regarded as one of the key bilateral issues with possible impact on peace and prosperity in the South Asian region [Moudgil (2015), Jha (2015)].

Both countries have set up different control structures for utilizing Teesta's water in their territories. The most significant structure in India is a barrage located at Gajaldoba in West Bengal, upstream of the international border. At this location, the flow is bifurcated into two channels: a diversion canal (Teesta Mahananda Link Canal or TMLC) and the main stream flowing into Bangladesh (see Figure 1). The obstruction of the flow has led to the formation of a pond upstream of the barrage, which is much smaller than typical reservoirs associated with dams. The water level of the pond is held constant by operating the lock gates of the above two channels. Thus, the incoming upstream flow at any instant may be regarded as the sum of the flows passing through the above two channels. The Government of Bangladesh seeks to utilize the flow of the river through another barrage and canal system centered at Duani, downstream of the border. Greater flow through the international border would permit greater diversion of flow at Duani. Control of the flow at Gajaldoba is crucial to the sharing of the river water between India and Bangladesh since there is no further diversion structure between this barrage and the international border.

The flow of the Teesta through the main stream at Gajaldoba has been recorded since the beginning of the year 1993. The flow through the TMLC has also been recorded since the end of 1997, when the canal became operational. The flow, measured in cubic meters per second (cumec), is recorded several times a day and subsequently averaged over a cycle of approximately 10 days (i.e., the first 10 days

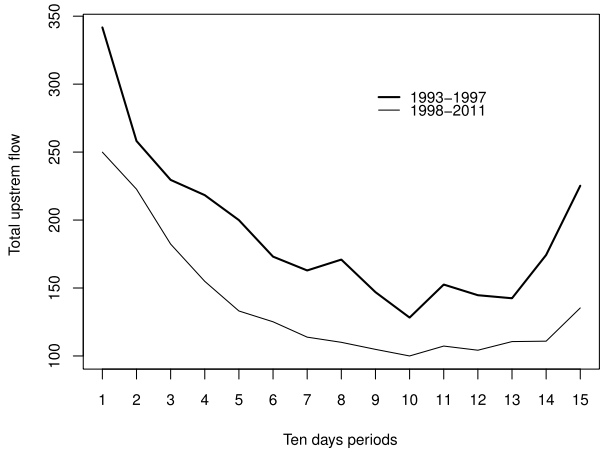

FIG. 2. *Plots of average upstream flow at the Teesta Barrage at Gajaldoba in different ten-day periods of the period November to March. The thick and thin curves correspond to the period January* 1993 *to March* 1997, *and the period January* 1998 *to December* 2011, *respectively.*

of a calendar month, the following 10 days and the remaining days of that month). In the sequel, we refer to these pre-determined periods as "ten-day periods." The work reported here is based on the flow data for the period 1993–2011 obtained by courtesy of the Department of Irrigation of the Government of West Bengal and the Central Water Commission of the Government of India.

The sum of the flows through the main stream and through the canal is regarded as the measured upstream flow. Figure 2 shows the average upstream flows during the periods 1993–1997 and 1998–2011 for the 15 ten-day cycles of the leanest months, namely, November to March. In order to maintain confidentiality of the quantum of actual flow, as required by the Ministry of Water Resources, Government of India, we have rescaled all the averages in such a way that the minimum value of the rescaled flow during 1998–2011 is 100 units. It is observed from Figure 2 that the average upstream flow in the time period 1998–2011 (thin curve) is much smaller than the corresponding flow in the period 1993–1997 (thick curve).

This large discrepancy cannot be explained by actual differences in flow during the two periods. The lean season flow is generated mostly by snow-melt and groundwater discharge in the upper reaches of the river, which is not very sensitive to year-to-year variation in the precipitation. However, there can be another explanation. The upstream flow during the period 1993–1997 consists entirely of the flow through the main stream, while a substantial part of the upstream flow for the period 1998–2011 comes from the flow through the TMLC. If, for some reason, there was any difference in the mechanisms of measurement of the two branches of flow, then that could be a reason behind this discrepancy.

At Gajaldoba, the flows through the two channels are measured indirectly. The total flow in a channel is regarded as the sum of flows through all the rectangular

gates in that channel. The discharge through a particular gate is computed from the *Villemonte formula* [Jain (2001), Subramanya (2013)]:

$$Q\left(h_1^{1.5} - h_2^{1.5}\right)^{0.385},$$

where h_1 and h_2 are the upstream and downstream water surface elevations, respectively, above the base of the gate. It is h_2 that is adjusted by operating the gate. The hydrological constant Q is computed from an empirically determined formula that depends on h_1, the dimensions of the gate and the acceleration due to gravity. The multiplier Q is different for different lock gates. In particular, one multiplier is used for all the main stream gates (of identical size) and another one is used for the canal gates (having identical dimensions but different from the main stream gates). These multipliers are empirically determined from controlled experiments prior to the installation of these gates at the site. Therefore, there is a possibility that the multipliers are not perfectly tuned to the operating environment. If this is the case, there would be an error in the measurements in the sense that all measurements would be distorted by an unknown multiplying factor, and this factor may not be the same for the main stream and the canal. The resulting error in the upstream flow, obtained by aggregation of the two measured flows, would depend on how much water is diverted through the canal. Diversion only began in 1998. This change, together with the distortions through multipliers, appears to be the reason behind the discrepancy observed in Figure 2.

Removal of this apparent distortion is important for proper implementation of decisions on water management, and also for eliminating any potential source of misunderstanding about that implementation. Such a misunderstanding is particularly undesirable in the context of the long history of dispute over sharing of the flow of this trans-boundary river [Menon (2015)].

A possible way of removing the distortion is to estimate each multiplier afresh through a controlled on-site experiment. In this experiment, the flows should be measured directly and compared with the estimates obtained through the existing indirect method. However, this experiment may involve much time and cost, and disrupt the regular task of managing the flow. An alternative solution based on existing data would be very useful.

It should be noted that, if one only has access to the possibly distorted measurements of flows through the two channels, then the two multipliers mentioned above are not individually identifiable. For the purpose of decision-making, however, knowing the correct ratio of the multipliers for the main stream and the canal flows is more important than knowing the multipliers themselves. The "absolute truth" corresponding to these measured flows may not be of much use unless they tally with the measurements made at the Bangladesh side of the border. Therefore, flows through both the channels at Gajaldoba need to be calibrated with measurements at the Bangladesh side. Once the flows through the two channels at the barrage at Gajaldoba are brought to a common scale of measurement, one can calibrate that common scale with the measurement of upstream flow at Duani, the

barrage at Bangladesh. This latter exercise can be done by using any one of the several methods of calibration of paired measurements that exist in the literature. Therefore, we restrict ourselves to the problem of matching the measurements of flow through the canal gates and the main stream gates by estimating the ratio of multipliers from historical data.

In the next section, we develop a model for the diversion mechanism on the basis of a careful study of the nature and the origin of the data. However, we do not presume any functional form of the underlying distribution. This brings us to a territory where there is no existing method of inference. In Section 3, we introduce a class of estimators for the correction factor, establish its consistency regardless of the underlying distribution and propose diagnostic tests for the model assumption. In Section 4, we use two specific estimators belonging to the proposed class to adjust the measurements of the bifurcated flow of the Teesta river at the Gajaldoba Barrage. We provide some concluding remarks in Section 5. The proof of the theoretical results and simulation studies of the small sample performances of the diagnostic tests and the two estimators are available as online supplementary material [Jana, Sengupta and Rudra (2016)].

2. Model specification.

2.1. *Problem formulation.* Let us denote the upstream flow through the river, flow through the canal and flow through the main stream as Z, X and Y, respectively. The quantity Z is unobserved, while the flows X and Y, regarded as random variables, are measured with distortion, as follows:

$$\begin{aligned} Z &= X + Y, \\ X_M &= \theta_1 X, \\ Y_M &= \theta_2 Y, \end{aligned} \tag{2.1}$$

where X_M and Y_M are measured values of X and Y, respectively, and θ_1 and θ_2 are unknown but fixed positive parameters. In particular, we can write the actual flow of the canal (X) and its measurements (X_M) as

$$\begin{aligned} X &= \sum_j X_{(j)} = Q_X \sum_j \bigl(h_{1X(j)}^{1.5} - h_{2X(j)}^{1.5}\bigr)^{0.385}, \\ X_M &= \sum_j X_{M(j)} = \theta_1 Q_X \sum_j \bigl(h_{1X(j)}^{1.5} - h_{2X(j)}^{1.5}\bigr)^{0.385}, \end{aligned} \tag{2.2}$$

where Q_X is the "correct" multiplier for the gates of the canal, $\theta_1 Q_X$ is the multiplier used in the calculation of flow, $h_{1X(j)}$ and $h_{2X(j)}$ are the gate-specific water elevations at the upstream and downstream sides of the gates, and the sum is over

all the gates. Likewise, for the main stream, we have

$$
\begin{aligned}
Y &= \sum_j Y_{(j)} = Q_Y \sum_j \bigl(h_{1Y(j)}^{1.5} - h_{2Y(j)}^{1.5}\bigr)^{0.385}, \\
Y_M &= \sum_j Y_{M(j)} = \theta_2 Q_Y \sum_j \bigl(h_{1Y(j)}^{1.5} - h_{2Y(j)}^{1.5}\bigr)^{0.385},
\end{aligned} \tag{2.3}
$$

where Q_Y is the "correct" multiplier for the gates of the main stream, $\theta_2 Q_Y$ is the multiplier actually used, $h_{1Y(j)}$ and $h_{2Y(j)}$ are the gate-specific water elevations at the upstream and downstream sides of the main stream gates, and the sum is over all the gates. The values of the multipliers used in the equation of X_M and Y_M may be wrong, that is, θ_1 and θ_2 may not be equal to 1. The parameters θ_1 and θ_2 are not identifiable from the data on X_M and Y_M alone. For matching X_M and Y_M, we set our goal as estimating the ratio of the multipliers, $\theta_0 = \theta_1/\theta_2$. We call this ratio a "correction factor" since X_M/θ_0 and Y_M are in the same scale of measurement, both being distorted by the factor of θ_2 from their respective true values.

Devising a suitable correction factor for matching measurements of flow through two channels is a form of calibration. However, we deliberately avoid using this term in our problem because it is somewhat different from the problem commonly referred to as "calibration" in statistical literature. In the latter problem, which is also known as "inverse regression," one variable is regarded as a function of the other variable coupled with additive or multiplicative random error. The task is to predict the original variable from its function using paired measurements. See Osborne (1991) and Greenwell (2014) for overviews of the techniques for inverse regression. These methods are not applicable to the present problem, where the actual flows X and Y are not necessarily functions of one another even in an approximate sense.

In contrast with usual models for calibration, we do not use any case-specific random error (additive and multiplicative) in model (2.1). In the absence of paired data of X and X_M (or Y and Y_M), such an error model would be difficult to handle. Instead, we opt for a simple model that might be amenable to a tractable solution to the real problem at hand.

If one assumes a specific form of the bivariate distribution of X and Y, then the problem of estimating θ_0 reduces to a standard problem of parametric estimation. However, the multipliers θ_1 and θ_2 are confounded with the scale parameters of X and Y, respectively. Consequently, θ_0 is confounded with the ratio of the scale parameters of X and Y. Therefore, even if the distributional form is correctly chosen, the parameter θ_0 may not be identifiable. The model needs to be strengthened with some additional assumption if θ_0 is to be inferred from the data.

2.2. *A strengthened model.* The upstream flow of the Teesta river at Gajaldoba is governed mostly by natural processes. There are a number of small dams on the river and its tributaries in the upper reaches. These are exclusively meant for

generating hydro-power. The holding capacities of these dams are relatively small. The disruption of natural flow caused by these dams are expected to have negligible impact on the average flow calculated over a period of ten days at Gajaldoba. Therefore, it is reasonable to assume that the "upstream flow" (Z) in model (2.1) is a natural phenomenon.

The target for the ratio of the canal flow and the main stream flow (X_M/Y_M) is decided several days ahead of the actual diversion. (It emerges from discussion with the project engineers that the standard operating procedure is to set a target for X_M/Y_M, rather than that for X_M.) During the lean season, the basis for the decision is generally the requirement of water for irrigation for the particular ten-day period, and the expected availability of the upstream flow for the relevant time of the year as anticipated from historical data. During the monsoon season, the demand of water for irrigation is less, and there is considerable leakage of water into the canal. In the remaining part of the year, the requirement for irrigation is practically the only determining factor for the proportion of diversion. Therefore, for a given ten-day period during that part of the year, it would not be unrealistic to assume that the ratio X_M/Y_M (and, consequently, the proportion $U = X/Z$) is determined by human need independently of the upstream flow Z, which is determined by natural phenomena.

As explained in Section 1, there are 36 "ten-day periods" in each year. We stratify the data by these periods, so that each year contributes a data point within a stratum, and estimate parameters by using the data from the relevant strata, where the independence of X_M/Y_M and Z is likely to hold.

In view of these considerations, we now formulate a more detailed version of the basic model (2.1). Let the number of strata be m. Suppose, for $i = 1, \ldots, m$, the pairs $(X_{M_{ij}}, Y_{M_{ij}})$, $j = 1, \ldots, n_i$, denote the n_i observations from the ith stratum, governed by the model

$$\begin{aligned} X_{M_{ij}} &= \theta_1 X_{ij} = \theta_1 Z_{ij} U_{ij}, \\ Y_{M_{ij}} &= \theta_2 Y_{ij} = \theta_2 Z_{ij}(1 - U_{ij}), \qquad j = 1, \ldots, n_i, i = 1, \ldots, m, \end{aligned} \tag{2.4}$$

for unspecified positive constants θ_1 and θ_2, where $Z_{i1}, Z_{i2}, \ldots, Z_{in_i}$ are unobserved samples from the distribution F_i defined over $[0, \infty)$, $U_{i1}, U_{i2}, \ldots, U_{in_i}$ are unobserved samples from the distribution G_i defined over $[0, 1]$, $i = 1, \ldots, m$, and *all the sets of unobserved samples are independent of one another*. Here, Z_{ij} and U_{ij} are the upstream flow and the fraction of diversion, respectively, in the ith ten-day period of the jth year, $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. Their randomness, represented through the distributions F_i and G_i, arises from the variation of these quantities from year to year. The observables of the model, $(X_{M_{ij}}, Y_{M_{ij}})$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ are distorted versions of the actual flows through the canal and the main stream [(X_{ij} and Y_{ij}), respectively, for $j = 1, \ldots, n_i$, $i = 1, \ldots, m$], and θ_1 and θ_2 are the unknown parameters representing the respective distortion factors of flow measurements in the two channels. The parameter of interest is $\theta_0 = \theta_1/\theta_2$.

3. Methodology.

3.1. *Derivation of the estimator.* While it is possible to assume some parametric forms of F_i and G_i and estimate θ_0 along with the ensuing nuisance parameters, such additional assumptions can only restrict the scope of applicability of the model. Estimation under a presumed distributional model could lead to bias in the estimator of θ_0 in case the assumed distribution is inappropriate. Instead, we look for an estimator that does not have to depend on a distributional assumption. This may be achieved by exploiting the independence between Z and U in model (2.4). While the assumption of independence is justified from the context of the application, the validity of this assumption may be checked through appropriate diagnostic tests, which we discuss in Section 3.3.

It may be recalled that the principle of independence between two random variables has been used in the past as the basis of inference in other situations also. A prominent example is the problem of blind source separation, where one seeks to separate the contributions of different sources in multiple linear mixtures (e.g., identifying different audio sources from signals recorded through one or more microphones), by making use of the assumption that the sources are independent [Comon (1994), Yu, Hu and Xu (2014)]. The problem of whitening of noise in signal processing involves filtering with parameters chosen to ensure uncorrelatedness/independence of the output time series [Levy (2008)]. Uncorrelatedness between linear estimators and linear zero function is sought to be ensured in estimation of parameters in a linear model [Sengupta and Jammalamadaka (2003)]. In Principal Components Analysis (PCA) also, one seeks to identify a Principal component (PC) that is uncorrelated with the PCs already identified [Anderson (2003)]. Thus, independence/uncorrelatedness has been an important criterion for statistical decision-making.

In the present situation, we can use a summary measure of dependence, and estimate θ_0 by a number which brings that measure close to the value it takes in the case of independence. Let D_n be an empirical measure of dependence, computed from paired samples of size n. We assume that values of D_n close to zero indicate lack of dependence, and values far away from zero indicate strong dependence. We also assume that the measure does not change when either the first number or the second one in the paired data values are multiplied with a common positive scale factor. The two common distribution free measures of dependence, Spearman's rho and Kendall's tau (see Examples 1 and 2 below), satisfy these two properties. Suppose, for $\theta \in \Theta \subseteq \mathbb{R}^+$ and $i = 1, \ldots, m$, the quantity $d_{in_i}(\theta)$ denotes the value of D_{n_i} when the latter is computed by regarding $((V_{i1}^*(\theta), W_{i1}^*(\theta)), \ldots, (V_{in_i}^*(\theta), W_{in_i}^*(\theta)))$ as the underlying data, where

$$
\begin{aligned}
V_{ij}^*(\theta) &= X_{M_{ij}} + \theta Y_{M_{ij}} \quad \text{and} \\
W_{ij}^*(\theta) &= \frac{X_{M_{ij}}}{X_{M_{ij}} + \theta Y_{M_{ij}}}, \qquad j = 1(1)n_i, i = 1(1)m.
\end{aligned} \tag{3.1}
$$

Note that the assumed scale-invariance of D_n implies that the value of $d_{in_i}(\theta)$ would remain unchanged if one uses $((V_{i1}(\theta), W_{i1}(\theta)), \ldots, (V_{in_i}(\theta), W_{in_i}(\theta)))$ as the data, where

$$V_{ij}(\theta) = \frac{V_{ij}^*(\theta)}{\theta_1} = \frac{Z_{ij}}{\theta_0}\big(\theta_0 + (\theta - \theta_0)(1 - U_{ij})\big), \quad \text{and,}$$
$$W_{ij}(\theta) = W_{ij}^*(\theta) = \frac{\theta_0 U_{ij}}{\theta_0 + (\theta - \theta_0)(1 - U_{ij})}, \qquad j = 1(1)n_i, i = 1(1)m. \tag{3.2}$$

When $\theta = \theta_0$, the above "data" reduces to $((Z_{i1}, U_{i1}), \ldots, (Z_{in_i}, U_{in_i}))$, which are pairwise independent random variables. The next lemma shows that this simplification happens only if $\theta = \theta_0$.

LEMMA 3.1. *Let the pairs* $(V_{i1}(\theta), W_{i1}(\theta)), \ldots, (V_{in_i}(\theta), W_{in_i}(\theta))$ *be defined as in* (3.2) *and* $(Z_{i1}, U_{i1}), \ldots, (Z_{in_i}, U_{in_i})$ *be samples from a bivariate distribution* H_i *for* $i = 1, \ldots, m$. *If there exists* $\theta = \theta^*$ *for which* $V_{ij}(\theta^*)$ *and* $W_{ij}(\theta^*)$ *are independent for all* $i = 1, \ldots, m$, *then this* θ^* *is unique.*

Independence of the Z_{ij}'s and the U_{ij}'s implies that $d_{in_i}^2(\theta)$ should be small for values of θ near the true value θ_0. This property should hold for each stratum. Accordingly, we define an estimator of θ_0 as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{m} d_{in_i}^2(\theta). \tag{3.3}$$

Use of various measures of dependence in (3.3) would produce different estimators. This class of estimators can be studied together.

Now let us consider some examples. Assume that $\{(V_k, W_k) : k = 1, \ldots, n\}$ is a set of paired random samples of size n drawn from the joint distribution of (V, W).

EXAMPLE 1. Spearman's rank correlation coefficient, rho [Spearman (1904)], between V and W is defined as

$$D_{S_n}\big((V_1, W_1), \ldots, (V_n, W_n)\big) = 1 - \frac{6\sum_{k=1}^{n} S_k^2}{n(n^2 - 1)}, \tag{3.4}$$

where S_k is the difference between the ranks of V_k and W_k, $k = 1, \ldots, n$.

EXAMPLE 2. Kendall's rank correlation coefficient, tau [Kendall (1938)], between V and W is defined as

$$D_{K_n}\big((V_1, W_1), \ldots, (V_n, W_n)\big) = \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{S_{k,l}}{n(n-1)}, \tag{3.5}$$

where

$$S_{k,l} = \operatorname{sgn}(V_l - V_k)\operatorname{sgn}(W_l - W_k), \qquad k, l = 1, \ldots, n,$$

and

$$\operatorname{sgn}(u) = \begin{cases} -1, & \text{if } u < 0, \\ 0, & \text{if } u = 0, \\ 1, & \text{if } u > 0. \end{cases}$$

3.2. *Consistency of the estimator.* Let the observations $(X_{M_{ij}}, Y_{M_{ij}})$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ follow the model (2.4). For $i = 1, \ldots, m$, let $H_{i\theta}$ be the bivariate distribution of $V(\theta) = (Z/\theta_0)(\theta_0 + (\theta - \theta_0)(1 - U))$ and $W(\theta) = (\theta_0 U)/(\theta_0 + (\theta - \theta_0)(1 - U))$, where $Z \sim F_i$ and $U \sim G_i$ are independent. Let $\mathbb{H}$ be the space of all bivariate distribution functions that are continuous almost everywhere, equipped with the metric ρ induced by the supremum norm. Let $D : \mathbb{H} \to \mathbb{R}$ be a measure of dependence such that $D(H) = 0$ whenever H is the product of its marginal distributions. Let $D_n(\mathbf{Z}, \mathbf{U})$ be a sample version of $D(H)$ computed from the samples $(Z_1, U_1), \ldots, (Z_n, U_n)$ of H, the samples being represented through the vectors $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ and $\mathbf{U} = (U_1, \ldots, U_n)^T$. Let this function be scale invariant, that is, $D_n(a\mathbf{Z}, b\mathbf{U}) = D_n(\mathbf{Z}, \mathbf{U})$ for any positive a and b. With $\mathbf{Z}$ and $\mathbf{U}$ defined as above, let us denote by $\mathbf{V}(\theta)$ and $\mathbf{W}(\theta)$ the vectors with elements

$$\begin{aligned} V_i(\theta) &= \frac{Z_i}{\theta_0}\big(\theta_0 + (\theta - \theta_0)(1 - U_i)\big) \quad \text{and} \\ W_i(\theta) &= \frac{\theta_0 U_i}{\theta_0 + (\theta - \theta_0)(1 - U_i)}, \qquad i = 1(1)n. \end{aligned} \tag{3.6}$$

We now list a number of conditions for proving the consistency of $\hat{\theta}$.

A. The set Θ is compact.

B. The dependence measure satisfies the following conditions:
   (i) $D : \mathbb{H} \to \mathbb{R}$ is continuous with respect to the metric space $(\mathbb{H}, \rho)$.
   (ii) D is bounded, $D(H_{i\theta_0}) = 0$ and $D(H_{i\theta}) \neq 0$, $\forall \theta \in \Theta \backslash \{\theta_0\}$, for $i = 1, \ldots, m$.

C. The function $D_n : \mathbb{R}^{2n} \to \mathbb{R}$ satisfies the following conditions:
   (i) The statistic $D_n(\mathbf{Z}, \mathbf{U})$ can be written as $a_n D(H^{(n)}) - b_n$, where $H^{(n)}$ is the bivariate empirical distribution function based on the data $(\mathbf{Z}, \mathbf{U})$, and a_n and b_n are real sequences such that $a_n \to 1$ and $b_n \to 0$ as $n \to \infty$.
   (ii) For $i = 1, \ldots, m$, whenever $(\mathbf{Z},\mathbf{U})$ are samples from $H_{i\theta_0}$, there is B_n such that $B_n = O_p(1)$ and the following inequality holds for all $\theta, \theta' \in \Theta$, almost surely in $H_{i\theta_0}$:

$$\big|D_n\big(\mathbf{V}(\theta), \mathbf{W}(\theta)\big) - D_n\big(\mathbf{V}(\theta'), \mathbf{W}(\theta')\big)\big| \leq B_n|\theta - \theta'|.$$

D. For $i = 1, \ldots, m$, the distributions F_i and G_i are absolutely continuous with respect to the Lebesgue measure.

THEOREM 3.2. *Let the estimator $\hat{\theta}$ defined by (3.3) be based on data arising from the measurement model (2.4). Let Conditions* A, B, C *and* D *hold. Further, let*

$$\frac{n_i}{n} \to \xi_i, \qquad \text{for some } \xi_i \in (0,1), i = 1, \ldots, m, \tag{3.7}$$

as $n = \sum_{i=1}^{m} n_i \to \infty$. Then

$$\hat{\theta} \xrightarrow{P} \theta_0.$$

Theorem 3.2 establishes consistency of the general class of estimators (3.3) based on different measures of dependence, which satisfy Conditions B and C. By verifying these conditions for the two examples mentioned in Section 3.1, we establish in the Supplementary Material [Jana, Sengupta and Rudra (2016)] the consistency of the estimators

$$\hat{\theta}_S = \arg\min_{\theta \in \Theta} \sum_{i=1}^{m} D^2_{S_{n_i}}\big((V_{i1}(\theta), W_{i1}(\theta)), \ldots, (V_{in_i}(\theta), W_{in_i}(\theta))\big), \tag{3.8}$$

$$\hat{\theta}_K = \arg\min_{\theta \in \Theta} \sum_{i=1}^{m} D^2_{K_{n_i}}\big((V_{i1}(\theta), W_{i1}(\theta)), \ldots, (V_{in_i}(\theta), W_{in_i}(\theta))\big), \tag{3.9}$$

where $D_{S_{n_i}}$ and $D_{K_{n_i}}$ have the forms as in (3.4) and (3.5), respectively.

3.3. *Diagnostic test for the assumption of independence.* The proposed methodology depends crucially on the assumption of independence of the aggregate Z_{ij} and the proportion U_{ij} in model (2.4), both being unobservable. However, they can be written in terms of the observables X_{Mij} and Y_{Mij} as

$$Z_{ij} = V_{ij}(\theta_0),$$
$$U_{ij} = W_{ij}(\theta_0),$$

where $V_{ij}(\theta_0)$ and $W_{ij}(\theta_0)$ are as defined in (3.2). The violation of the assumption of independence of Z_{ij} and U_{ij} may occur in two ways:

H_1^A: there is no θ for which $V_{ij}(\theta)$ and $W_{ij}(\theta)$ are independent, and

H_1^B: there exists a θ^* such that $V_{ij}(\theta^*)$ and $W_{ij}(\theta^*)$ are independent, but $\theta^* \neq \theta_0$.

The scenario H_1^A means that the observables X_{Mij} and Y_{Mij} are such that it is not possible to find a θ that makes the pair of statistics $X_{Mij} + \theta Y_{Mij}$ and $X_{Mij}/(X_{Mij} + \theta Y_{Mij})$ independent. Therefore, these observables cannot possibly be synthesized through model (2.4) with independent Z_{ij} and U_{ij}.

The scenario H_1^B means that there is a θ^* such that $X_{Mij} + \theta^* Y_{Mij}$ (or $X_{Mij}/\theta^* + Y_{Mij}$) and $X_{Mij}/(X_{Mij} + \theta^* Y_{Mij})$ are independent, but the distributions of $(X_{M_{ij}}/\theta^* + Y_{M_{ij}})$ and $(X_{M_{ij}}/\theta_0 + Y_{M_{ij}})$ are not identical. Here,

$(X_{M_{ij}}/\theta_0 + Y_{M_{ij}})$ would be the apparent quantum of flow of the river as measured through the lock gates of the main stream if the entire flow is allowed to pass through the main stream. There is record of this flow for a few years when the canal had not been operational. The scenario H_1^B would then mean that, with θ^* chosen to make $(X_{Mij}/\theta^* + Y_{Mij})$ and $X_{Mij}/(X_{Mij} + \theta^* Y_{Mij})$ independent, the distribution of $(X_{M_{ij}}/\theta^* + Y_{M_{ij}})$ is not what it should be, as known from historical records of an appropriate time.

The scenarios H_1^A and H_1^B may be regarded as alternative hypotheses, where the corresponding null hypotheses are

H_0^A: there exists a θ^* such that $V_{ij}(\theta^*)$ and $W_{ij}(\theta^*)$ are independent,
H_0^B: $V_{ij}(\theta_0)$ and $W_{ij}(\theta_0)$ are independent (i.e., $\theta^* = \theta_0$).

Note that the hypotheses H_0^B and H_1^B constitute a partition of H_0^A. In other words, H_0^B is nested in H_0^A. Since these hypotheses are to be tested for a diagnostic purpose, it is their nonrejection that would indicate the appropriateness of the presumed model (so that the proposed estimator makes sense). It transpires that the hypothesis H_0^A has to be tested first. If it cannot be rejected, one has to proceed by assuming that it is true. The hypothesis H_0^A implies, through Lemma 3.1, that either H_0^B is true or H_1^B is true for a unique θ^*. If H_0^B cannot be rejected in favor of H_1^B, then one is left with the assumption that H_0^B is true. The latter hypothesis is equivalent to the assumption of independence between Z_{ij} and U_{ij}.

We would provide a testing procedure for each of the testing problems mentioned above.

Test of H_0^A vs. H_1^A. The null hypothesis H_0^A can be tested against the alternative H_1^A by checking the independence between $V_{ij}(\theta^*)$ and $W_{ij}(\theta^*)$, where $\theta^* = \hat{\theta}$, the minimizer of a chosen squared measure of dependence (e.g., $\hat{\theta}_S$ or $\hat{\theta}_K$). Any standard nonparametric test for independence [see Hájek, Šidák and Sen (1999), page 126] may be adapted to this problem as follows. For example, for $i = 1, \ldots, m$, let ρ_i be Spearman's rho statistic for the data $(V_{i1}(\hat{\theta}_S), W_{i1}(\hat{\theta}_S)), \ldots, (V_{in_i}(\hat{\theta}_S), W_{in_i}(\hat{\theta}_S))$, where $\hat{\theta}_S$ is as defined in (3.8). Let $T_i = \rho_i/\sqrt{\mathrm{Var}(\rho_i)}$, where $\mathrm{Var}(\rho_i) = \frac{1}{(n_i-1)}$, be the standardized version of the statistic under H_0^A. For testing H_0^A against the general alternative H_1^A, we use the pooled statistic $T = \frac{1}{\sqrt{m}} \sum_1^m T_i$. The asymptotic null distribution of this statistic should be standard normal. As another example, we can replace ρ_i by the Kendall's tau statistic τ_i computed from the data $(V_{i1}(\hat{\theta}_K), W_{i1}(\hat{\theta}_K)), \ldots, (V_{in_i}(\hat{\theta}_K), W_{in_i}(\hat{\theta}_K))$, where $\hat{\theta}_K$ is as defined in (3.9), and scale the statistic by the square-root of $V(\tau_i) = \frac{2(2n_i+5)}{9n_i(n_i-1)}$. The pooled test statistic would have the same form and the same distribution as above [Kendall (1938)].

Test of H_0^B vs. H_1^B. As noted earlier, if the hypothesis H_0^A is assumed to be true and θ^* is the value of the parameter that makes $(X_{M_{ij}}/\theta^* + Y_{M_{ij}})$ and $X_{M_{ij}}/(X_{M_{ij}} + \theta^* Y_{M_{ij}})$ independent, then the hypothesis H_0^B is equivalent to the equality of the distributions of $(X_{M_{ij}}/\theta^* + Y_{M_{ij}})$ and $(X_{M_{ij}}/\theta_0 + Y_{M_{ij}})$. The latter quantity represents the total flow if all of it is measured through the lock gates of the main stream. The total flow had indeed been passed through the main stream and measured during the period 1993–1997 (when only the main stream gates of the river had been operational). We denote this *benchmark* data by $Y'_{M_{i1}}, \ldots, Y'_{M_{in'_i}}$, $i = 1, \ldots, m$, where n'_i is the number of years during the above period for which main stream flow data for the ith ten-day cycle are available. For every i, each of the quantities $Y'_{M_{i1}}, \ldots, Y'_{M_{in'_i}}$ have the same distribution as $(X_{M_{ij}}/\theta_0 + Y_{M_{ij}})$. As for the distribution of $(X_{M_{ij}}/\theta^* + Y_{M_{ij}})$, one can estimate θ^* by the value of θ that minimizes a squared empirical measure of dependence (e.g., $\hat{\theta}_S$ and $\hat{\theta}_K$). Therefore, testing for H_0^B reduces to a set of two sample problems—one for each stratum. Specifically, if $\hat{\theta}_S$ is used, then the first sample for the ith stratum would consist of $(X_{M_{i1}}/\hat{\theta}_S + Y_{M_{i1}}), \ldots, (X_{M_{in_i}}/\hat{\theta}_S + Y_{M_{in_i}})$, while the second sample in the same stratum would be $Y'_{M_{i1}}, \ldots, Y'_{M_{i,n'_i}}$. For each stratum, one can use the Wilcoxon signed rank test after standardizing it appropriately [Hollander and Wolfe (1999), page 108]. In order to make the test usable for modest sample size, we take the sum of these standardized test statistics across the m strata, and standardize the sum by the pooled standard deviation, $\sqrt{m}$. This pooled and standardized test statistic should be approximately standard normal under H_0^B.

A preliminary simulation study of small sample properties of this test statistic (not reported here) revealed no problem with normality but a bit of excessive variability. This problem may be solved by a scale adjustment to the test statistic through Bootstrap resampling. Specifically, the resampling may be done independently within each stratum by drawing random paired samples with replacement from the available paired data, the sample size being the same as that of the original data. Resampling from the benchmark data set may be done similarly. The sample standard deviation of the test statistics obtained from these resamples can then be used to standardize the test statistic obtained from the original data.

In view of the nested nature of the two null hypotheses, the test of H_0^A should be followed by the test of H_0^B, and rejection at either stage would indicate inappropriateness of the model (2.4). Further, in accordance with the Bonferroni inequality, probabilities of $\alpha/2$ may be allocated to each of the tests in order to achieve the overall significance level α for the combination of two tests [Miller (1981)].

3.4. *Simulation of performance.* A simulation study of the performance of the tests was conducted by generating data from parametric models representing the alternative hypotheses H_1^A and H_1^B. The results show that the power of the nested

tests, for sample size comparable to the data analyzed in the next section, is adequate whenever the empirical bias of the estimator is large. Additional simulations show that the efficiencies of the estimators $\hat{\theta}_S$ and $\hat{\theta}_K$ are somewhat insensitive to the number of strata when the sample size per stratum is held fixed, but improve with decreased number of strata when the total sample size is held fixed. Another simulation study reveals near-unbiasedness of a bootstrap estimator of standard error obtained by independent resampling with replacement from each stratum. The details of the studies are available in the online supplementary material [Jana, Sengupta and Rudra (2016)].

4. Analysis of Teesta river flow data. As mentioned in Section 1, the data collected at the Gajaldoba barrage consists of measurements of flow through the main stream (Y_M) during the period 1993 to 2011, averaged over ten-day periods, and similar data for flow through the canal (X_M) during 1998 to 2011.

There are thirty-six ten-day periods in a calendar year, as explained in Section 1, which may be treated as different strata. As explained in Section 2, we have to exclude from our analysis the strata corresponding to the rainy season (June to September) and the lean season. The driest of the ten-day periods happens to be the first ten-day period of February when the average aggregate flow (i.e., the sum of observed flows through the canal and the main stream) reaches its minimum value. We exclude all the strata for which the average aggregate flow is less than 150% of this minimum level. We also exclude the first ten-day period of October since the average aggregate flow in that stratum is more than the flow during the first ten-day period of the rainy season. These considerations leave us with 12 strata: three ten-day periods from each of the months of April, May and November, the last two ten-day periods of October and the first ten-day period of December. It is expected that, during these periods, the proportion of diversion had been decided independently of the amount of upstream flow. We would check this assumption empirically also.

The main data set consists of measurements of canal flow and main stream flow during the twelve chosen ten-day periods of the calendar years 1998–2011. Thus, there are up to 14 observations per stratum, though there are a few missing observations. For the purpose of model diagnostics, we also use the data for the above strata during the calendar years 1993–1997 when the canal was not operational and measurements of flow through the main stream represented the total flow. For this period, the sample size per stratum was 5, except for one stratum, where a single observation was missing.

The problem of assessing a correct multiplier for adjusting the two sets of flow measurements amounts to estimation of the parameter θ_0 from these data under the measurement model (2.4).

We begin the analysis by performing the diagnostic tests proposed in Section 3.3 on the above data in order to check the validity of the model (2.4). Rescaling of the second test statistic (see Section 3.3) is done on the basis of 500 bootstrap samples,

TABLE 1
Estimates of θ_0 by using (3.8) and (3.9), together with the bootstrap estimate of the standard error and bootstrap confidence interval

	$\hat{\theta}_S$	$\hat{\theta}_K$
Estimates	0.76	0.76
Bootstrap S.E.	0.0713	0.0729
Bootstrap 95% C.I.	(0.64, 0.92)	(0.63, 0.91)

as in the simulations. The p-values for the test of H_0^A against H_1^A, corresponding to the estimators (3.8) and (3.9), happen to be 0.21 and 0.31, respectively. The p-values for the test of H_0^B against H_1^B, corresponding to the estimators (3.8) and (3.9), turn out to be 0.73 and 0.71. These finding indicate the validity of the model (2.4) for the Teesta riverflow data. We shall later present the result of an additional graphical check.

The values of the proposed estimators (3.8) and (3.9) are reported in Table 1. The table also shows standard errors and 95% confidence intervals, computed from nonparametric bootstrap estimates based on 1000 resamples. Here the bootstrapping for bivariate data is done within each stratum with replacement, the sample size being the same as the stratum size.

The two estimates have similar values, indicating that a measured flow of one cumec through the main stream gates is about 76% of a measured flow of one cumec through the canal gates. Further, the bootstrap confidence interval corresponding to each estimate excludes the value 1. Thus, the correction factor is significantly smaller than 1. This conclusion is important in the context of decision-making. Unless the measurements from the canal and the main stream gates are brought to a common scale, there would never be a proper control over the flow.

The dashed curve in Figure 3 is the plot of the adjusted upstream flow during the lean season of 1998–2011 [i.e., sum of flow through the main stream (Y_M) and the corrected flow though the canal ($X_M/\hat{\theta}$)] obtained by using the estimators (3.8) or (3.9). This curve is close to the average upstream flow in the time period 1993–1997 (solid thick curve). This finding also indicates that the two estimators under the chosen model, which provide for calibration through a single multiplier, lead to reasonable adjustment.

5. Concluding remarks. In this article, we have introduced a class of nonparametric estimators for the purpose of correcting the bifurcated components of an aggregate flow, assuming that there is a true correction factor. The method developed here may apply generally to other situations where an aggregate quantity is split into two channels and then measured in different ways, necessitating correc-

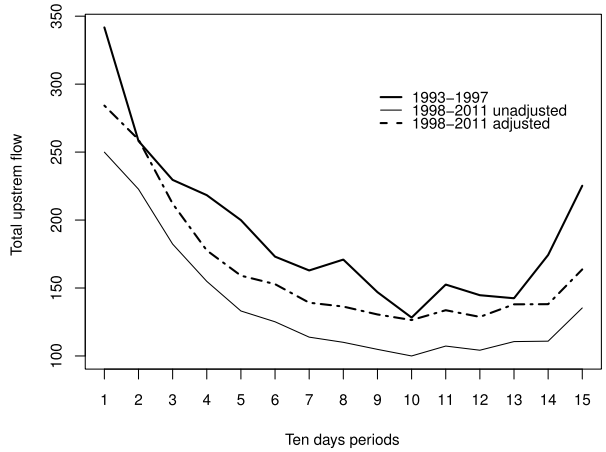

FIG. 3. *Plots of average upstream flow at the Teesta Barrage at Gajaldoba in different ten-day periods of the period November to March. The solid thick curve corresponds to the period January* 1993 *to March* 1997. *The solid thin curve corresponds to the unadjusted flow during the period January* 1998 *to December* 2011. *The dashed curve is the adjusted flow for the latter period, obtained by using estimators* (3.8) *or* (3.9).

tion of the measurements of the parts. For example, in diagnostic and therapeutic medicine, a ray of light is used to determine the optical properties of living tissue. The tissue surface receives short pulses of light (Z) emitted from a small source, through a laser beam or optical fiber. Amounts of absorption (X) and scattering (Y) are observed over a period of time for studying the relationship between absorption and scattering coefficients of the tissue [Cheong, Prahl and Welch (1990), Patterson, Chance and Wilson (1989)]. The amounts of absorption and scattering are estimated by different approximation formulae so that their estimated values, X_M and Y_M, may differ from X and Y, respectively. Correction may be needed here [Section 4.2 of Wilson and Patterson (2008)]. The proportion of light absorbed (X/Z) is a property of the tissue, while the amount of incident light (Z) depends on external factors such as intensity and distance of the light source. These two quantities should be independent, as long as the frequency of light does not change. Another instance of bifurcation arises while determining light absorption in a system of particles for characterizing different transmission media, such as highly scattering particles, colloids and composite materials, by using measurements of transmittance and reflectance [see Duncle and Bevans (1956) and Tassan and Ferrari (2002)].

Apart from the two measures considered here (Kendall's tau and Spearman's rho), one could possibly use other well-known measures of dependence such as the product moment correlation and distance correlation [Székely, Rizzo and Bakirov (2007)]. However, consistency of the corresponding estimator is not guaranteed.

In particular, Condition B(i) is not satisfied by the product moment correlation, and we could not prove Condition C(ii) in the case of distance correlation. The estimates of the correction factor based on these two measures of dependence, for the data set on Teesta flow through the canal and the main stream at Gajaldoba, happen to be 0.79 and 0.75, respectively. These numbers are in line with the finding of Section 2 of the online supplementary materials.

Since the discrepancy among the different point estimates of the correction factor obtained through the estimators (3.8) and (3.9) is small, either one can be used. The simulation study shows that the two estimators have comparable performance. For the data at hand, the 95% confidence interval corresponding to each estimate excludes the value 1. This situation calls for action. The government may consider the costs and benefits of conducting a controlled experiment in case the need for better accuracy is felt. The present analysis, including Figure 3, shows that adjustment through either of the proposed methods would be a better option than not adjusting at all.

The simple methods of correction presented in this article may be extended to the situation where an additive quantity is multifurcated into more than two components, which are measured with different degrees of fixed and multiplicative distortion. This type of problem may arise in analyzing trifurcated flow of a river through a barrage with two canals. A case in point is the recent operationalization of a new canal (Teesta Jaldhaka Main Canal or TJMC) at the Teesta Barrage at Gajaldoba. This research problem may be considered in the future.

Acknowledgments. The inference problem considered in this paper arose from an assignment received by the third author from the Government of West Bengal in November 2011. The data used for the analysis of Section 4 was received in the context of that assignment from the Department of Irrigation, Government of West Bengal, with permission from the Central Water Commission, Government of India. The authors thank the two agencies for providing the data. The authors also thank the two referees, the Associate Editor and the Editor for their critical comments that led to substantive improvement of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Correction of bifurcated river flow measurements from historical data: Paving the way for the Teesta water sharing treaty" (DOI: 10.1214/16-AOAS958SUPP; .pdf). Section 1 of the Supplementary Material contains technical proofs of the theoretical results. Section 2 describes results of the simulation. Section 3 shows the computation of the Cramer Rao Lower Bound used in these simulations.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

CHEONG, W. F., PRAHL, S. A. and WELCH, A. J. (1990). A review of the optical properties of biological tissues. *IEEE J. Quantum Electron*. **26** 2166–2185.

COMON, P. (1994). Independent component analysis, a new concept? *Signal Process*. **36** 287–314.

DUNCLE, R. V. and BEVANS, J. T. (1956). An approximate analysis of the solar reflectance and tranmittance of a snow cover. *J. Meteor.* **13** 212–216.

GREENWELL, B. M. (2014). *Topics in Statistical Calibration*. Ph.D. thesis, Air Force Institute of Technology. MR3218115

HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. *Probability and Mathematical Statistics*. Academic Press, San Diego, CA. MR1680991

HOLLANDER, M. and WOLFE, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York. MR1666064

INDIA TODAY (2015). Mamata Banerjee raises Teesta issue with Sheikh Hasina, assures a breakthrough. *India Today*, Feb 25.

JAIN, S. C. (2001). *Open-Channel Flow*. Wiley, New York.

JANA, K., SENGUPTA, D. and RUDRA, K. (2016). Supplement to "Correction of bifurcated river flow measurements from historical data: Paving the way for the Teesta water sharing treaty." DOI:10.1214/16-AOAS958SUPP.

JHA, R. K. (2015). India-Bangladesh politics over Teesta river water sharing. *South Asia Monitor*, Jan 27.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

LEVY, B. C. (2008). *Principles of Signal Detection and Parameter Estimation*, 1st ed. Springer, New York.

MENON, M. S. (2015). Time to look at Teesta. *The Indian Express*, Aug 13.

MILLER, R. G. JR. (1981). *Simultaneous Statistical Inference*, 2nd ed. Springer, New York. MR0612319

MOUDGIL, M. (2015). South Asian water wars: An improbability. *World Policy Insti.*, Sep 14. Available at http://www.worldpolicy.org/blog/2015/09/14/south-asian-water-wars-improbability.

OSBORNE, C. (1991). Statistical calibration: A review. *Int. Stat. Rev.* **59** 309–336.

PATTERSON, M. S., CHANCE, B. and WILSON, B. C. (1989). Time resolved reflectance and transmittance for the non-invasive measurement of tissue optical properties. *Appl. Opt*. **28** 2331–2336.

RUDRA, K. (2012). *Atlas of Changing River Courses in West Bengal*. Sea Explorers' Institute, Kolkata.

SENGUPTA, D. and JAMMALAMADAKA, S. R. (2003). *Linear Models*: *An Integrated Approach*. *Series on Multivariate Analysis* **6**. World Scientific, River Edge, NJ. MR1993512

SOLOMON, S. (2010). *Water*: *The Epic Struggle for Wealth*, *Power and Civilization*. Harper Collins, New York.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol*. **15** 72–101.

SUBRAMANYA, K. (2013). *Engineering Hydrology*, 4th ed. Tata McGraw Hill Education, New Delhi.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist*. **35** 2769–2794. MR2382665

TASSAN, S. and FERRARI, G. M. (2002). A sensitivity analysis of the transmittance–reflectance method for measuring light absorption by aquatic particles. *J. Plankton Res*. **24** 757–774.

VIDAL, J. (2010). How water raises the political temperature between countries. *The Guardian*, June 25.

WILSON, B. C. and PATTERSON, M. S. (2008). The physics, biophysics and technology of photodynamic therapy. *Phys. Med. Biol*. **53** 61–106.

YU, X., HU, X. and XU, J. (2014). *Blind Source Separation*: *Theory and Applications*, 1st ed. Wiley, New York.

K. JANA
D. SENGUPTA
APPLIED STATISTICS UNIT
INDIAN STATISTICAL INSTITUTE
KOLKATA 700108
WEST BENGAL
INDIA
E-MAIL: kaushikjana11@gmail.com
sdebasis@isical.ac.in

K. RUDRA
WEST BENGAL POLLUTION CONTROL BOARD
PARIBESH BHAVAN
10A, BLOCK LA
SECTOR III, SALTLAKE CITY
KOLKATA 700098
WEST BENGAL
INDIA
E-MAIL: rudra.kalyan@gmail.com