Coupling of modular linear and nonlinear regression models, devising a unit multilinear regression model for flow prediction at Alaknanda HEP intake.

P. MACHHKHAND

GHD Pty Ltd, City Plaza Centre, PO BOX 14352, Doha, Qatar
pradyumna.machhkhand@ghd.com.au

Abstract A multilinear regression model is built from the inferences of modular linear and nonlinear regression models to predict the flow of Alaknanda River at its hydroelectric project intake station, Alaknanda HEP. The data consists of 10-daily flows at stations Joshimath and Lambagarh for a recorded length of 33 years and the catchment area. From a deep prognosis, it is concluded that the flow data of Lambagarh (D/S) should be transferred to Alaknanda HEP(U/S). The correlations of Lambagarh discharge values are found to possess a poor correlation against Joshimath flow data. Since Joshimath flow data is an approved data, it enhances the study to check the seasonal correlation. The objective of the study is to predict the missing values of Lambagarh discharge by applying modular models technique followed by coupling the developed modular linear and non-linear models by optimization. The concept of modularity is 'divide and conquer'. The modularity conveys the differential arrangement of flows with respect to seasons. It preserves the over estimation and under estimation of flow. The seasonal analyses of the data are run through correlation and regression. An algorithm is set to induce the intercorrelation analyses among the seasons wherever the seasonal correlations themselves are not satisfactory. Sequentially, applying regression, the flow values are predicted. The parameters most sensitive to these predicted values of flows are optimized by an optimum threshold value and the discharges lying below this threshold are treated with percentage trough evaluation of flow hydrograph corresponding to the season for which the intercorrelation with other seasons have been of due importance. Alternatively, when the seasonal intercorrelation fails, the partial series of the investigated season is chosen and subjected to regression. The chosen series is validated by iteration till the t-stat value is within the confidence band. The modular output discharges are then assembled and subjected to multi linear regression model as input. The modelled results are post processed through the aforesaid optimization technique, by estimating the percentage of flow values in the hydrograph trough, to derive the final flow hydrograph of Lambagarh, which is finally transposed to Alaknanda HEP on a catchment area ratio basis.

Key words intercorrelation; modular models; multiregression; percentage trough evaluation; threshold; t-stat; HEP

INTRODUCTION AND BACKGROUND INFORMATION

The hydroelectric project of 300 MWAlaknanda HEP lies in the state of Uttranchal, India. The River Alaknanda, which along with Bhagirathi River and other tributaries constitutes River Ganga, originates in glacial regions of the Himalayas in the extreme northern parts of the district of Chamoli in Uttaranchal. The control structure is a dam proposed at the intake site, Alaknanda HEP is situated on the river Alaknanda. At this site the catchment area is 1010 Sq Km. Lambagarh and Joshimath sites are located downstream of Alaknanda HEP with catchment areas of 1200 Sq Km and 4508 Sq Km respectively.

In hydropower projects, the study of water availability is always a matter of utmost importance to attain the flow duration curves for estimation of power potential.

Therefore, the time series of discharge data used in these projects are always crucial. In the present study, the 10-daily flows at stations Joshimath and Lambagarh for a recorded length of 33 years are available to derive the flow series at Alaknanda HEP. However, the flow series of Lambagarh are available from UP Irrigation department of India with a wide range of data gap that are required to be filled in. Although the flow series at Joshimath are approved, the same would not be feasible to transfer it to the upstream catchment i.e., Alaknanda HEP because the data is recorded at G&D site downstream of the confluence of rivers, Dhauliganga and Alaknanda whereas, Lambagarh is the only station with available data at the downstream of the intake. Due to the reason stated above, the availability of Lambagarh 10daily flow series data is selected for transposing it to Alaknanda. However, the 10daily data is statistically tested against Joshimath data, a priori considered as authentic.

The objective of the study is to predict the missing values of Lambagarh discharge by applying modular models technique based on linear and nonlinear regression and to calibrate the results by identifying the parameters to be optimized.

MODULAR AND MULTI-REGRESSION MODELS

In the arena of modelling, especially time series data, global models are preferred to any other for obtaining results and their assessment. Although the approach is quite significant, the results may appear erroneous unless the parameters most sensitive to the models are optimized. The discharge hydrograph is widely distributed over time with undulated crests and toughs of varying magnitudes. While modelling time series data per se of greater length, the output gets entrapped locally between the peak flows and the low flows. More precisely, it is commonly observed that when the model is capable of capturing the peak, it may be falling short to capture the low flows and vice-versa. Ample methods and models are available to overcome this phenomenon of global models and the goal is to form a global model with minimized errors.

Application of modular models is also a technique to deal with the aforesaid problems. Modular models are also termed as local models. The basic concept of modularity is ‘divide and conquer’ (Jain *et al.,* 2002). The modularity conveys the differential arrangements of data with respect to class. The input data is divided into classes and each class forms a model. The inferences from each model are combined to form a series of data. These data are then subjected to regression and this chain of regression models is termed as multi-regression model.

In the present context, the modularity is represented in the model by classifying the available time series flow data of Joshimath and Lambagarh on a seasonal basis, which are treated as input. It is viable to divide the data on a seasonal basis because different seasons have effects on the observed flow playing a significant role in mapping the catchment in terms of climatological influences therein.

The divided models, termed as local models, comprising a particular season provide good check for any discrepancies in the data because the same may be overlooked while treating the model as global.

In the present study, the hydrological cycle is divided into four seasons for framing into models. The seasons are identified as monsoon, non-monsoon, snow accumulation and snowmelt. From the observed data of both the stations, Lambagarh and Joshimath, it is found that there exist a wide gap of data in the Lambagarh flow series and the flow variations in the snow accumulation season appear to be unscrupulous and this is also evidenced by the inferences from correlation analysis of Lambagarh data versus

Joshimath data.

DATA ORGANIZATION

The seasonal modular models are named as, SM1, SM2, SM3, SM4 representing the seasons viz., monsoon, non-monsoon, snow accumulation and snowmelt respectively.

The seasonal classified data comprising both Lambagarh and Joshimath are plotted below Fig.1:

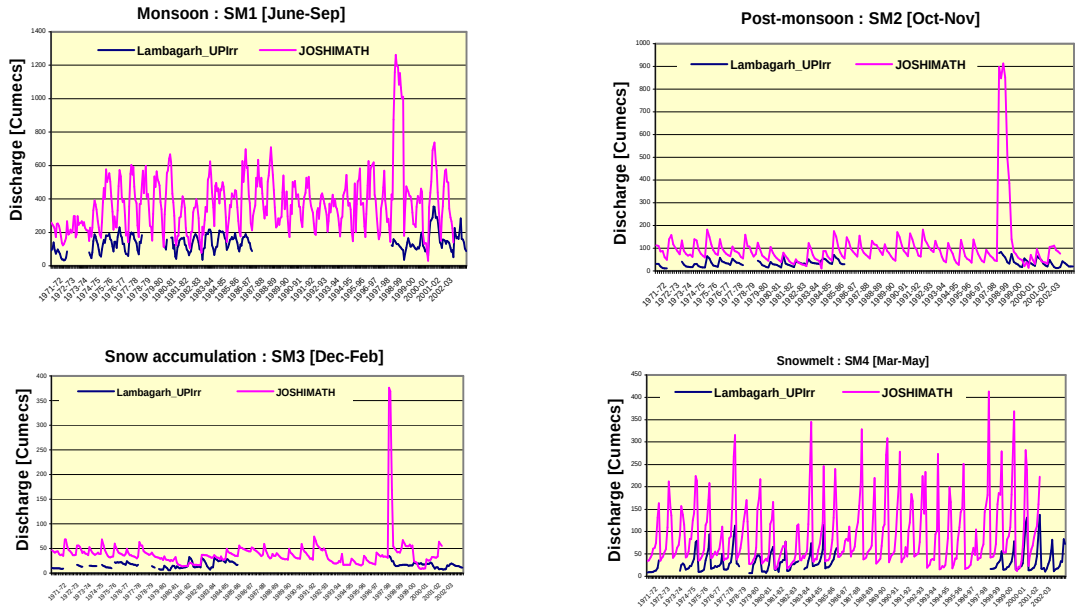

Fig. 1 Plots of modular models input data for the classified seasons

For each model, the correlation analysis is carried out to check whether the data is useful for its application in regression models and if not then, the improvement of the same is completed by treating the data as explained in the next section.

Correlation is the degree to which two or more quantities are linearly associated. In a two-dimensional plot, the degree of correlation between the values on the two axes is quantified by the so-called correlation coefficient expressed as, $\rho\ (X,Y)$ where X and Y are two different data sets (Haan, 1977), say, X as effective rainfall and Y as discharge independent of units.

Mathematically, it is expressed as,

$$\rho\,(X,Y) = \mathrm{cov}\,(X,Y)/\sigma_x\sigma_y \tag{1}$$

where $\mathrm{cov}\,(X,Y)$ is the covariance between X and Y and correspondingly, σ_x *and* σ_y are the standard deviations. The expressions are as follows:

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{2}$$

$$\sigma_x = \sqrt{\left(\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 \Big/ n \right)} \quad , \ \overline{X} \text{ is the mean of X} \tag{3}$$

$$\sigma_x = \sqrt{\left(\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2 \Big/ n \right)} \quad , \ \overline{Y} \text{ is the mean of Y} \tag{4}$$

Equation (1) can be re-written as,

$$\rho\,(X, Y) \quad = \quad \sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right) \Big/ \sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 \sum_{i-1}^{n} \left(Y_i - \overline{Y} \right)^2}$$

The correlation co-efficient, r is plotted below in Fig.2 for all the four seasons with variables, *Y* as Lambagarh data from UP Irrigation and *X* as Joshimath discharge series:

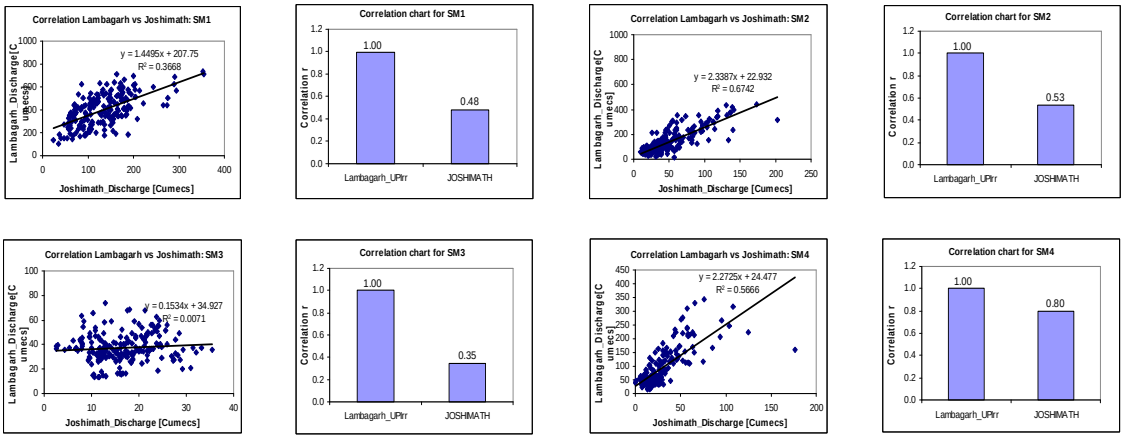

Fig. 2 Correlation chart of the classified seasonal models

The results are also tabulated sequentially in Table 1 and remarks are drawn based on rough rule of thumb (Larose, 2006):

Table 1 Correlation co-efficient (r) for classified seasonal models

Seasons	ID	Correlation co-efficient, r	Remarks
Monsoon	SM1	0.48	Variables are mildly positively correlated
Non - monsoon	SM2	0.53	Variables are mildly positively correlated
Snow accumulation	SM3	0.35	Variables are mildly positively correlated
Snowmelt	SM4	0.80	Variables are positively correlated

Although the r-value of snow accumulation period (SM3) is 0.35, it is close to threshold value of 0.33 and is presumably, unfit for regression analysis. Therefore, the

inter-correlation is carried out between Y-variable of SM3 as static and X-variable of SM1, SM2 and SM4. Consequently, the results show very poor correlation and the inter-correlation approach is abandoned.

However, using the same seasonal data of SM3, the degree of freedom is altered by mapping the data sequentially and it is observed that the data of this season is useful only with a degree of freedom not exceeding 26 which fosters to build two regression models with option-1 (df = 26) and option-2 (df = n-1) where n is the total number of observations made for the season. The correlation co-efficient, r for option-1 is computed as 0.65. It is also noted that the continua have broken after three consecutive hydrological years for the season SM3 and the data onwards appeared to be erroneous because the flow values have exceeded the Joshimath flows which is practically not possible as it is in the downstream and the data are recorded after the confluence of rivers, Dhauliganga and Alaknanda at Joshimath. The reason for the same might be manual fuzziness of data. In the regression analysis, option-1 is adopted and the learning process is supported by optimization as explained elaborately in Section-5 and Section-6

MODELLING SCHEME

The scheme of the model is framed below as shown in Fig.3 for simplicity. It also describes the stages involved in modelling from input of the model to calibration.

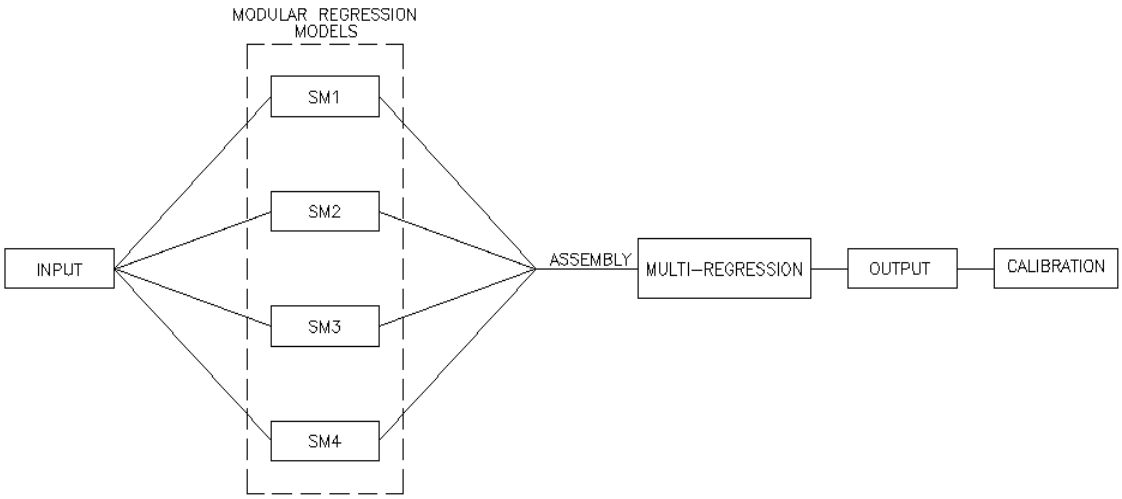

Fig. 3 Modelling scheme of modular models

The scheme is mathematically expressed as,

$$F(x)_i = (\beta_0)_i + \beta_i x \tag{5}$$

where i = [1..4] corresponding to SM1, SM2, SM3, SM4 respectively

Upon assembly, the predicted values are derived as,

$$F_1(x) = \beta_0' + \beta' x \tag{6}$$

where $F_1(x)$ is further subjected to calibration.

The derived relationship of Equation (6) is an outcome of multiple applications of regression and therefore, it is termed as multi-regression model, which would, understandably, necessitate calibration of results to reduce the errors.

MODEL APPLICATION

The pre-processing of data described in section-3 assures the quality of data to be used as inputs while applying the model for prediction of Lambagarh flow. The seasonal regression equations are formed on modular basis as shown in equation (5).

While $F(x)_i$ being a function representing the predicted values of Lambagarh data for each season, the other parameters, $(\beta_0)_i$ and β_i are intercepts and the gradients respectively for $i = [1..4]$. These parameters are derived from the standard least square method (Larose, 2005) using the observed flow data of Lambagarh and Joshimath, the target being Lambagarh flow.

Ideally, the combination of $F(x)_i$ where $i = [1..4]$ should represent a function with intercept, β_0' and gradient, β' as unique. Therefore, the outcome of function $F_1(x)$ as expressed in equation (6) is none other than the regression applied over the assembled $F(x)_i$ and Joshimath as input.

The input data is subjected to the test for stability of the mean and is widely known as t-Test. The same data is split up into two different series randomly chosen without affecting the order of the data set. This test is based on null hypothesis that when the two data series aforesaid are normally distributed then, the difference between the mean values of the two series is equal to zero. Assuming this to be true, the test statistic has a Student's t distribution (Haan, 1977).

The subsets the data are named as, X_1 and X_2 and are tested for stability of mean.
Taking $\overline{X_1}$ and $\overline{X_2}$ as the averages, the test statistic is given by,

$$t = \left(\overline{X_1} - \overline{X_2}\right) \Bigg/ \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)^{0.5} \tag{7}$$

where n_i is the number of data in the subset
$\overline{X_i}$ is the mean of the subset X_i
s_i^2 is the variance of the subset X_i

From the Student's- t distribution, the critical region with 5% significance level is set as,

$$\{-\infty, t(\upsilon, 2.5\%)\} U \{t(\upsilon, 97.5\%), +\infty\} \tag{8}$$

If a data set is under the aforesaid hypothetical condition, signifying no presence of

outliers in the series then, any result of t-Test that falls out of the confidence band as expressed in equation (8) disapproves the null hypothesis. However, in the current research, the t-stat value is within the confidence band and thus, it is verified to be modelled. The assembled seasonal models are displayed below in Fig.4

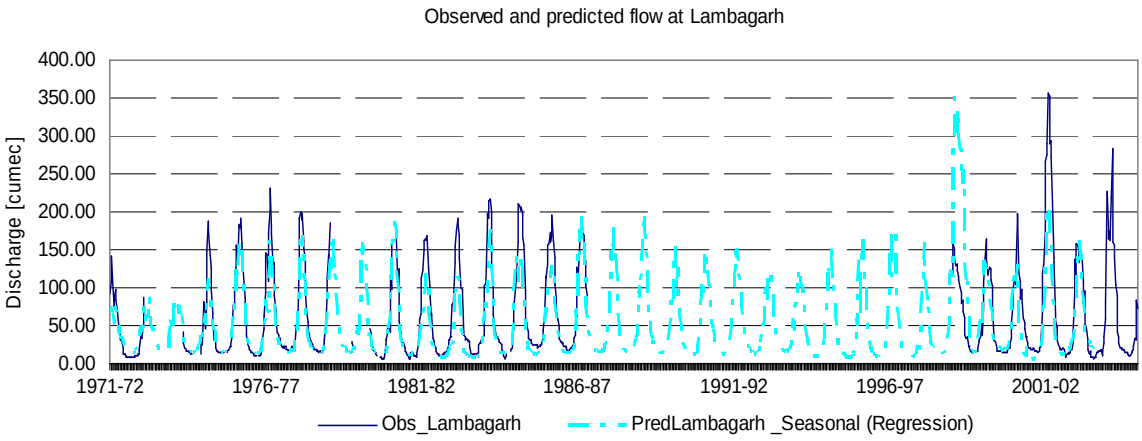

Fig. 4 Predicted flow of Lambagarh from assembled seasonal model

IDENTIFICATION OF PARAMETERS FOR OPTIMIZATION

The result obtained above fosters the need to calibrate the model because as an impact of assembly, it neither meets the objective where the peak values are concerned nor the values in the trough are estimated with accuracy. This enhances the research to identify the parameters to be optimized so that the low flow values are well-targeted vis-à-vis the peak flows may not be under-estimated. It is, indeed, essential here to find a constraint that would restrict the low flows to be overestimated or underestimated while optimizing the parameters (Press *et al.*, 2002). The importance of low flow values is due to the fact that model SM3 dictates immense errors and it lies within the period of lean season flows i.e., October to May defined for Alaknanda HEP showing a consistent trend over these periods. The lean season flows are very important to estimate the design flood for intake structure and barrage for various construction periods.

The parameters intercept, β_0' and gradient β' have significant effect in the linear multi-regression model. Any iteration on either of the parameters would change the result substantially. In the present case, the predicted values are modelled well in terms of rising and falling limbs of the hydrographs but they end up with erroneous results at the troughs. In fact, the low flows are overestimated. This indicates that, even though both the parameters are highly sensitive to the prediction, at least the parameter gradient β' may be kept constant while optimizing vis-à-vis the intercept β_0' should be investigated upon fixing the threshold value for low flows.

In hydroelectric projects, the water availability study for power potential estimation is equally important as the flood estimation for design of dams. From the flow duration curve, the 90% discharge is utilized for design of spillways and these discharges are categorized as low flows. Therefore, while optimizing the aforementioned parameters, the criticality of low flows is of due importance.

PERCENTAGE TROUGH CRITERION

The percentage trough comprising low flow values are set by a threshold value derived from the exceedance probability (Rao *et al.*, 2000) plot of annual lowest flow series as shown in Fig.5. The plot shows curves with both observed and predicted Lambagarh low flow values corresponding to a significant probability percent. The point of intersection of these two curves indicates that the flow values, lying in the curves, less than the magnitude observed at intersection point should be analyzed and evaluated. The quantitative evaluation of these numbers, in the present research, is termed as Percentage Trough Evaluation. In it, initially, the point of intersection in the probability plot is taken as a threshold and the percent of values less than this threshold in the flow hydrograph of observed and predicted Lambagarh are defined as the percentage trough flows.

The percent obtained for the predicted flow is equated with observed flow of Lambagarh for same threshold by optimizing the parameter β_0' of equation (6) and preserving the gradient β'.

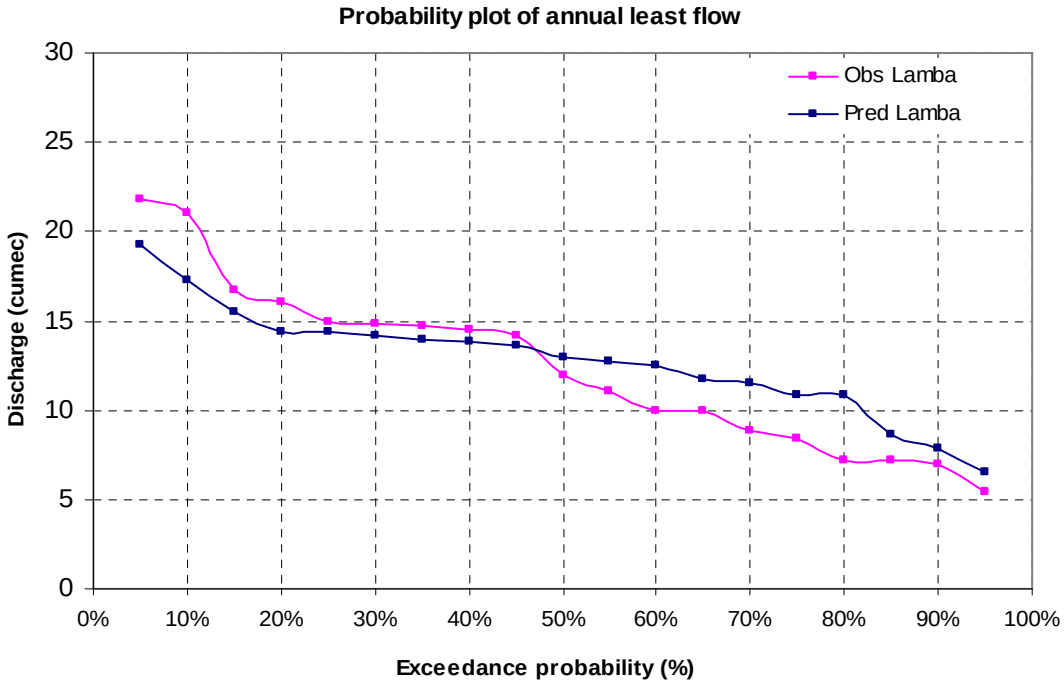

Fig. 5 Exceedance probability plot of annual least flow

PARAMETER OPTIMIZATION

The percentage criteria can be expressed mathematically as follows:

Let γ_{obs} = % *of values in the obsereved flow hydrograph* $\leq$ *Threshold*

γ_{pred} = % *of values in the predicted flow hydrograph* $\leq$ *Threshold*

Putting a constraint, $\gamma_{pred} = \gamma_{obs}$, the objective function is given by,

$$f(\gamma_{pred}) = w\beta_0^{'} \tag{9}$$

where *w* is weight associated with the intercept $\beta_0^{'}$ of equation (6)
The weight *w* is introduced to illustrate that γ_{pred} is directly proportional to $\beta_0^{'}$.

The parameters *w* and $\beta_0^{'}$ are simultaneously taken into account to optimize the function and it may be re-expressed as,

$$F_1(x) = w\beta_0{'} + \beta' x \tag{10}$$

or, $G(x) = \varepsilon + \beta' x$
(11)
where $G(x) = F_1(x)$ *and* $\varepsilon = w\beta_0^{'}$

The optimization is exercised for three different threshold values extracted from the probability plot and the reasons for the same are explained below in the following section.

RESULTS AND DISCUSSIONS

The three threshold values adopted to calibrate the model are:

i. The point of intersection or the point closest to it on the probability plot i.e., 13.63
ii. Point showing approximate constant values, consecutively, in the predicted curve i.e., 11.72
iii. Point showing approximate constant values, consecutively, in the predicted curve i.e., 10.82

Since the errors are not distributed proportionally, it is least likely that reducing the errors, while considering the threshold at the point of intersection, would lead to a symmetric closure of the curves and therefore, the sensitivity of the model is checked by selecting the other two points randomly as explained above. Albeit the points chosen are pseudo-hypothetical, nevertheless, the likelihood of occurrence of point could not be ignored at the chosen threshold. It enhances the study of error estimation from model results and the same is discussed below.

The root mean square error is given by,

$$RMSE = \sqrt{\left(\sum_{i=1}^{n} (P_i - O_i)^2 \Big/ n \right)} \tag{12}$$

where *P* and *O* are predicted and observed values respectively.
n is the number of data in the sample.

The results obtained for the parameter ε from all the three thresholds are listed below in Table 2.

Table 2 Optimized parameter (ε) for different threshold values with respective errors

Threshold	ε	RMSE
10.82	3.710	41.92
11.72	3.760	41.90
13.63	4.669	41.69

From the estimation of errors, the root mean square error, RMSE is lowest for the threshold 13.63 and is most suitable to adopt for functional optimization. As a result of calibration, the new series of flow hydrograph is plotted.

The RMSE of the model, Model 1, at the combination unit, which is merely the assembly of seasonal regression models is computed and compared with the model, Model 2, derived from the regression of these seasonal models, which are outcome of regression at all distinguished seasons. Since the later model is a product of multi-regression analysis, it is termed, in this study, as multi-regression model. The comparative analysis of errors in terms of RMSE is listed below in Table 3.

Table 3 Root mean squared errors for pre and post optimized models

Models with predicted flow at Lambagarh	RMSE
Model 1 (Assembly of seasonal regression models)	59.33
Model 2 (Post optimized multi-regression model)	41.69

Therefore, Model 2 is adopted for filling in the data gaps of observed Lambagarh flow and then the entire flow series is transferred to Alaknanda intake by catchment area ratio. The ratio is given by,

$$\left\{Q_{Alaknanda} / Q_{Lambagarh}\right\} = \left\{A_{Alaknanda} / A_{Lambagarh}\right\} \tag{13}$$

The observed flow at Lambagarh, the post optimized model with predicted values of flow at Lambagarh and the transposed flow series at Alaknanda HEP intake are shown below in Fig.6 .The scale of Fig. 6 is changed to magnify the low flow values and the same is clearly shown in Fig. 7

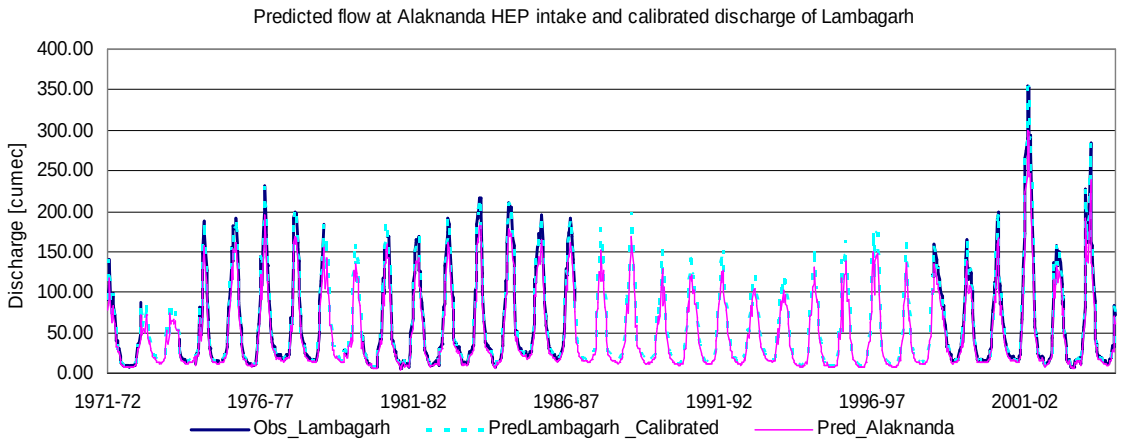

Fig. 6 Predicted Lambagarh after calibration and predicted flow at Alaknanda

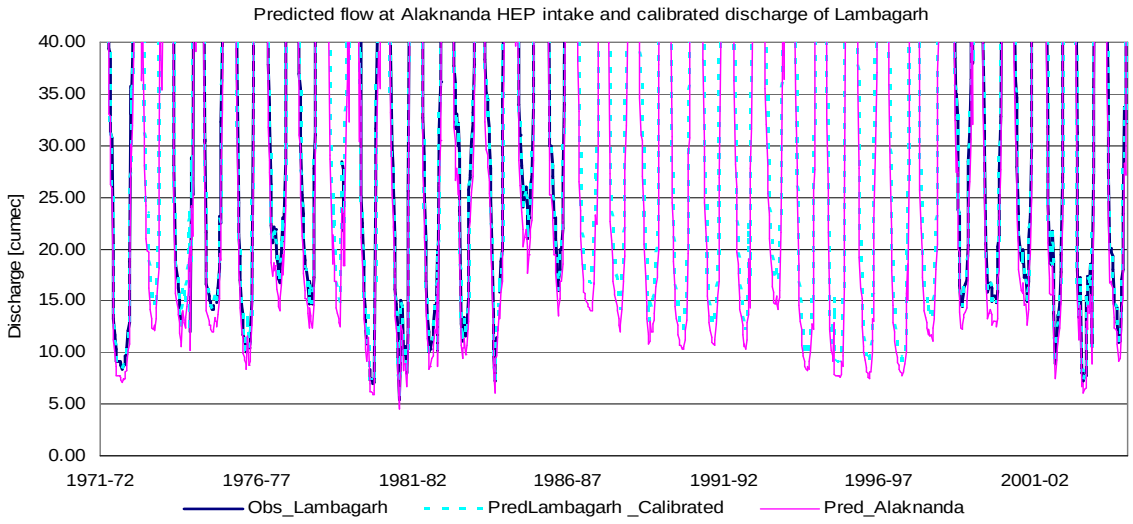

Fig. 7 Predicted Lambagarh and predicted flow at Alaknanda with modified scale

SUMMARY AND CONCLUSIONS

The result obtained is significant and noticeable. Though the method involved in both developing the model and calibrating it is quite rigorous, it covers the physical responses of flow at Lambagarh and then by understanding whether the responses are correlated with the D/S flow at Joshimath have been encapsulated in modular form. This differentiates the study from a conventional linear relationship. However, once the modular models based on linear and non-linear regression are formed, the assembled unit of these models is basically a raw model of seasonal, which is further processed

through regression and a unit regression model is developed. This unit model is linearly defined and the parameters are identified for calibration of model.

The method adopted for optimizing the parameter duly considers the low flow values. However, there is no strict distinction between low flows and high flows as it is beyond the scope of this study. Moreover, the lean season flows are more important in any hydroelectric project and therefore, the objective is focused more on these flows. In other words, the probability exceedance plot of annual lowest flow series curves is indicative of the fact that if a plane is introduced through point of intersection of both the curves, it may form an axis that divides the curves symmetrically or nearly symmetric. It may be inferred from the same that any parameter that is identified in the curves for optimization would also compensate for the errors observed in high and peak flows. Though a number of multi-objective algorithms are available for optimization, the adopted method suits the project needs and it may be applicable where a combination unit is expected to produce errors.

Acknowledgements This work would have been due without the support of Ian Burton and David Birch. It is recognized and made presentable with the support of Howard Graham, GHD Pty Ltd, Doha, Qatar. The kind guidance of Robin D Connolly and Ross Fryar is acknowledged and highly appreciated.

REFERENCES

L. C. Jain., J. Kacprzyk (2002). Studies in fuzziness and soft computing: New learning paradigms in soft computing, Springer, 103-110.

Charles T. Haan., (1977). Statistical methods in hydrology, 1st edition, The Iowa State University Press, 50-250.

Daniel T. Larose., (2006). Data mining methods and models, Wiley, Hoboken, New Jersy, 45-150.

Daniel T. Larose., (2005). Discovering knowledge in data: An introduction to Data mining, Wiley, Hoboken, New Jersy.

William H. Press., William T. Vetterling., Saul A.Teukolsky., Brian P.Flannery., (2002). Numerical recipes in C: The art of scientific computing, 2nd edition, Cambridge University Press, 394-396

A R. Rao., Khaled H.Hamed (2000). Flood frequency analysis, CRC Press LCC, Boca Raton, Florida, 1-20.