TU Delft | Deltas, Infrastructures & Mobility Initiative

SmartPhones 4 Water

Understanding headwater baseflow contributions to the overall water supply of the Kathmandu Valley, Nepal

W.S. Brouwer[1], A. van Hamel[1], M. van Haren[1], P.E. Kindermann[1], R.P. Verboeket[1], Dr. ir. J.C. Davids[3] and Dr. T. Bogaard[2]

[1]Delft University of Technology, Master programme Civil Engineering
[2]Delft University of Technology, Department of Watermanagement
[3]California State University, Chico, Department of Civil Engineering and College of Agriculture

Abstract. The Kathmandu Valley in Nepal is facing the combined effects of population growth, rapid urbanization, economic development, and climate change. This results in serious water management challenges: growing freshwater demands, declining water tables, drying of streams, and deteriorating water quality. Insufficient surface water supplies have led to increased reliance on groundwater, especially during the dry winter and pre-monsoon seasons (November - May). Despite groundwater's importance, it is sparsely measured, poorly understood, and insufficiently managed. As it is difficult and costly to measure all groundwater extractions in the Valley, a water balance approach is an alternative method to estimate total net groundwater pumping. Therefore, the aim of this research was to develop and evaluate potential methods for quantifying total pre-monsoon baseflow supplies by extrapolating baseflow measurements of a subsample of watersheds to unmeasured watersheds. Estimated baseflow was used, together with other water balance fluxes and changes in storage, to evaluate net groundwater pumping in the Valley. Three different methods were used: (1) Spatial Analysis, (2) Regression Model, and (3) Black Box (machine learning). All methods relied on streamflow data from 2017 to 2019, collected by citizen scientists from S4W-Nepal. Based on the three methods we presented, we cautiously conclude that it is possible to determine the pre-monsoon baseflow contributions from a sub-sample of head water catchments. Total baseflow estimates for the Valley using Spatial Analysis, Regression Model, Black Box were 2.32, 2.30, 2.65 m^3/s respectively. These values show orders of magnitude that correspond with expected values. By using the average baseflow values of all three methods, we were able to close the water balance and make an assumption for the net groundwater pumping in the Valley. Based on a population of 3.5 million, a net groundwater extraction of 96 L/person/day during pre-monsoon was found. This striking outcome emphasizes the need for more discharge and groundwater extraction measurements, to decrease the uncertainties and to refine the methods.

Keywords: Baseflow calculation, Hydrology, Kathmandu Valley, Nepal, Machine Learning, Regression Model, Spatial Analysis

I. Introduction

LOCATED in the foothills of the Himalayas, the Kathmandu Valley in Nepal (Valley) is home to more than 2.5 million permanent residents (population in 2011) [1]. With a population growth of 4% per year, the Valley is one of the fastest-growing metropolitan areas in South Asia [2]. For the period 2011-2031 population growth is expected to be 52% [1]. The combined effects of population growth, rapid urbanization, economic development, and climate change are resulting in serious water management challenges.

Growing demand for freshwater on one side, and declining water quality and quantity on the other side, put pressure on existing water resources.Surface water supply were long

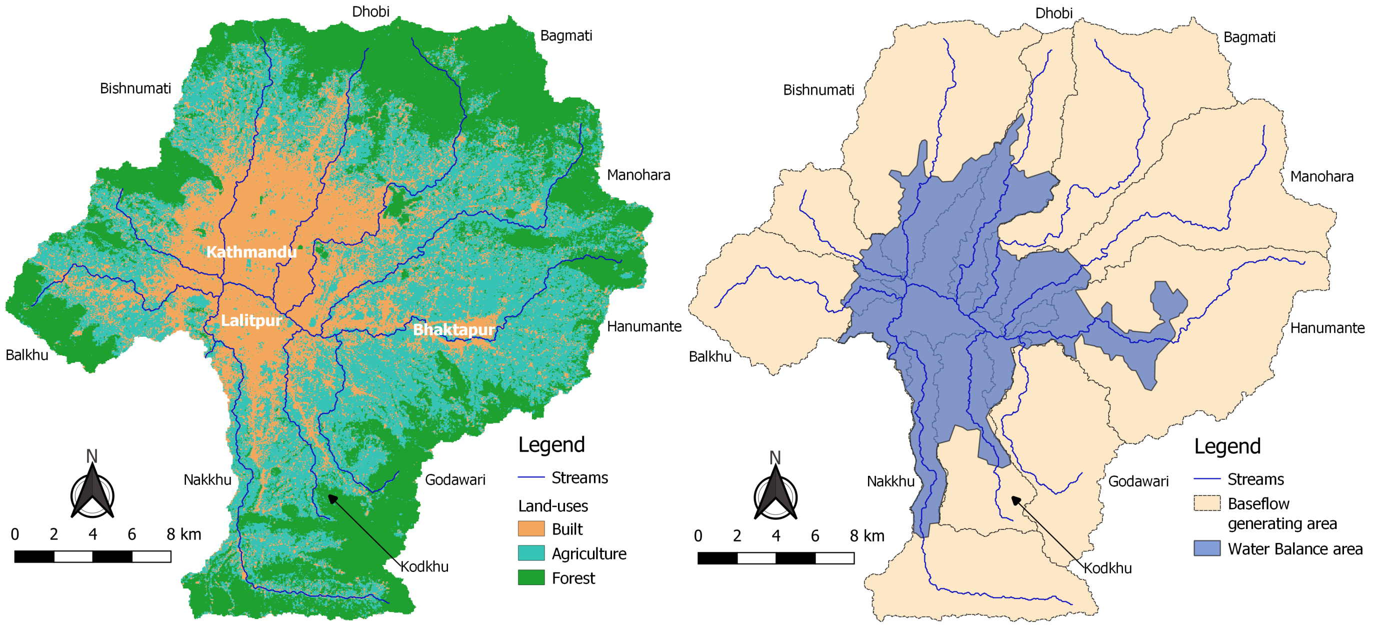

Fig. 1. 1.1: Locations of the rivers in the Kathmandu Valley and the three land-use classes, 1.2: Baseflow generating (tan) and water balance (blue) areas. Net groundwater pumping is solved for only the blue water balance area.

outpaced and now the Valley depends on groundwater as the main water supply, especially during the winter and pre-monsoon season (November - May), when only 20% of the annual precipitation occurs [3]. Annually, groundwater extractions exceed recharge rates, resulting in groundwater level declines [4]. The public agency responsible for water supply to the Kathmandu Valley, the Kathmandu Upatyaka Khanepani Limited (KUKL), can only meet 19% and 31% of the total water demand (estimated to be 370 million litres per day) for the dry and wet seasons, respectively [5]. The deficit between demand and supply is currently filled by other sources such as stone spouts, springs, rainwater harvesting and extractions from privately owned wells [6] [7] [8]. According to Udmale et al., the deficit between water demand and supply will increase to 322 million litres per day by 2021 [9]. As groundwater is considered to be a clean water source, groundwater is increasingly relied up within the Valley [10]. As a result, the groundwater levels in the shallow unconfined aquifer dropped from 2.57 to 21.58 meters below ground surface between 2003 and 2014. Moreover, the monsoon is an essential source of groundwater recharge in the Valley. However, ungoing hardscaping (e.g. paving) of the Valley's landscape, and changes to monsoon timing and duration due to climate change will also impact groundwater recharge in the Valley [11].

Groundwater is discharged in two ways: (1) by using natural springs and baseflow in streams or (2) by artificially removing the water from wells. As baseflows and spring supply are insufficient during the winter and pre-monsoon months, manual groundwater extractions must make up the difference. Therefore, it is important to quantify the amount of groundwater extraction. However, it is difficult to measure all places where water is manually extracted in the Valley. A potential method of estimating total net groundwater pumping in the Valley is applying a surface layer water balance. Unfortunately, baseflow contributions to the Valleys water supply are poorly measured. Seeing this gap, Smartphones4Water-Nepal (S4W-Nepal) started measuring discharge from several head water catchments throughout the Valley in 2017. However, there are multiple drainage areas that are not captured by these measurements. In this paper, we investigated methods to solve this problem.

A. *Research question*

In this research we focused on improving the understanding of baseflow contributions to the Valley's pre-monsoon water supply. Therefore, the aim was to develop and evaluate potential methods for extrapolating baseflow measurements to unmeasured catchments, to develop a robust estimate of the total baseflow contribution to the Valleys pre-monsoon water supply. The estimated baseflow will be used, together with other water balance fluxes and changes in storage, to evaluate net groundwater pumping in the Valley.

To obtain net groundwater pumping, the following questions had to be answered:

1) How can monthly stream flow measurements from a subsample of headwater catchments in the Valley be used to determine total pre-monsoon baseflow contributions to the Valley's water supply?
2) Can these estimates of pre-monsoon baseflow be used in a water balance to estimate net groundwater pumping?

II. Study Area

The Kathmandu valley is located at an elevation of approximately 1,400 m and covers an area of around 587 km^2 [4].

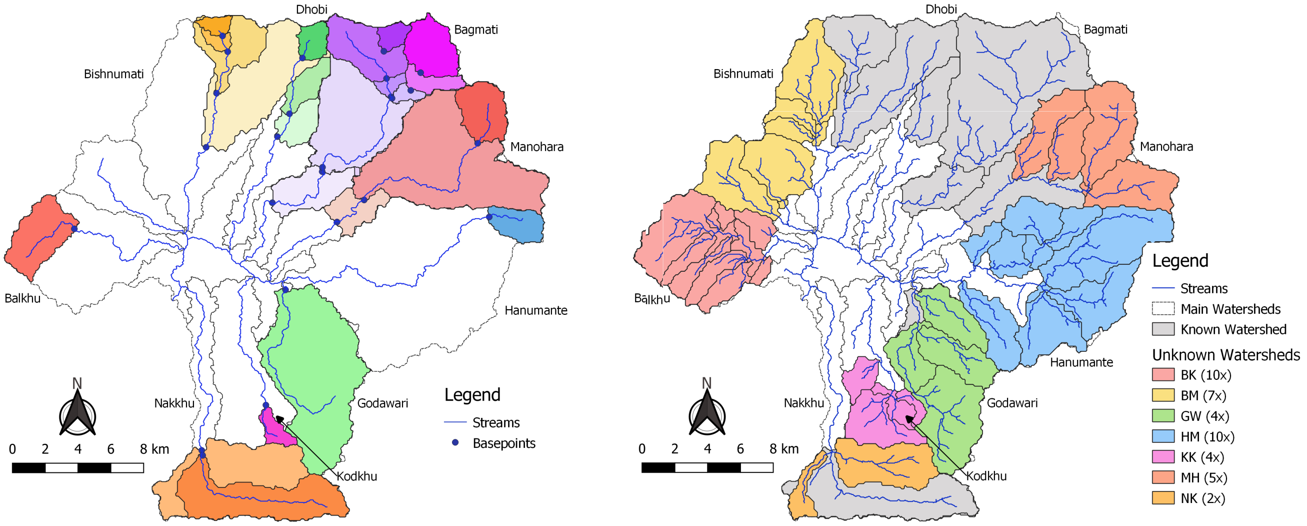

Fig. 2. 2.1: Locations of the basepoints in the Kathmandu Valley and the corresponding known watersheds, 2.2: Locations of the unknown watersheds within the Valley.

With a mild slope and surrounded by steep hills and mountains, the Valley has always been an attractive place for humans to settle. The Valley is principally drained by the Bagmati River, whose headwaters originate from the Northern region of the Valley. Eight other tributaries (Balkhu, Bishnumati, Dhobi, Godawari, Hanumante, Kodku, Manohara and Nakkhu) join the Bagmati river, prior to exiting at the southwestern edge of the Valley near Chobar. From upstream to downstream land-use changes from mainly natural to built in the Kathmandu, Lalitpur, and Bhaktapur urban areas (Figure 1.1). The Valley and its surrounding hills consist of old basement rock which is covered with unconsolidated and consolidated sediments [12]. The impermeable rock layer is bowl shaped and the only water outlet can be found at Chobar, where the Bagmati river leaves the Valley in the south-west direction. Water can enter the shallow aquifer especially in the northern part of the Valley. Elsewhere, the Valley's floor is mainly covered by an impermeable clay layer which separates the shallow aquifer from the deep aquifer. Recharge of the deep aquifer is possible, but predominantly in the northern part of the Valley [13]. Natural recharge of the shallow aquifer is declining due to the increasing sealing of the surface by urbanisation which prohibits infiltration. Currently, groundwater is extracted from both the shallow and deep aquifer. To estimate net groundwater pumping in the Valley we defined the spatial domain of our water balance. We decided to select the densely urbanised area where the land-use is predominantly completely built. The water balance area is visualised in Figure 1.2.

III. Methodology

To be able to answer the research questions, we split up the methodology in several smaller steps. Firstly, we analysed discharge data to determine the baseflow $[m^3/s]$ during pre-monsoon for different baseflow measurement locations throughout the Valley, so-called basepoints. For the basepoints, corresponding watersheds were delineated, and specific baseflows were computed [L/km^2]. Secondly, we investigated different characteristics (e.g. precipitation and land-use) to determine their correlation to the specific baseflow. Thirdly, we constructed three different methods to predict baseflow for unknown watersheds. The three different methods that were used are:

1) Spatial Analysis
2) Regression Model (manual)
3) Black Box (machine learning model)

Applying the three methods resulted in different estimations of baseflow for the unknown watersheds. We compared the three methods to discuss their strengths and weaknesses. By comparing the results, the overall surface water inflow of the water balance could be determined. Then, we estimated the other fluxes and changes in storage. Finally, we applied these estimates to close the water balance and to estimate net groundwater pumping in the Kathmandu Valley.

A. *Definition of baseflow for known basepoints and watersheds*

For the period 2017-2019 citizen scientists and members of S4W-Nepal have collected discharge data at different locations around the Valley by performing USGS mid-section measurements with a SonTek FlowTracker Acoustic Doppler Velocimeter (ADV). With the collected data, we were able to plot and analyse hydrographs for these locations. Since we found out that the number and regularity of measurements per location was highly variable, we decided to only use locations with more than two measurements during the pre-monsoon to determine baseflow. This resulted in 25 basepoints spread over the Valley for which the baseflow could be deducted based on hydrographs. Since we were working with limited data, we decided to use a three year average for the pre-monsoon months since there were no noticeable differences in the baseflow values for the different years. To normalise for catchment size, we converted the baseflow [m^3/s] to the specific baseflow [L/s/km^2].

Based on the 25 basepoints, we delineated the 25 corresponding *known watersheds* using Quantum Geographic Information System (QGIS) (see Figure 2.1). The known watersheds are located all around the Valley, but most densely in the northeast. Watersheds for which the baseflow was not measured are called *unknown watersheds* within this paper. The locations of the unknown watersheds can be found in Figure 2.2.

B. *Evaluation of Characteristics impacting baseflow*

Specific baseflow from the 25 known watersheds was not the same. Therefore, we compiled the variables that would likely influence the baseflow for each watershed. The constructed Regression Model and Black Box used these to build and train their algorithms, but logically these models needed the right variables as input data. For this paper, these variables are referred to as *characteristics*. We hypothesized that the following nine characteristics might impact specific baseflow of a watershed:

1) Precipitation (P)
2) Land-use (3 classes: natural, agricultural and built)
3) Basepoint elevation
4) Evaporation (ET)
5) Stream orientation
6) Area (A)
7) Stream length
8) Mean slope
9) Presence of shallow aquifer/recharge areas

We obtained the following the watershed characteristics using QGIS: area, stream orientation, basepoint elevation, mean slope, and streamlength (total length of all streams within the watershed). Also, the presence of the shallow aquifer and the recharge areas, expressed in percentages of the total area, were obtained from QGIS. For evaporation, USGS Simplified Surface Energy Balance Operational (SSEBop) remotely sensed data was used in combination with QGIS. For land-use classification we made use of the study on ”Quantifying the connections” from Environmental Monitoring and Assessment (EMAS) by Davids [4]. The six original land-use types were merged to three classes: natural (forest and shrubs), agricultural (rice and non-rice), and built (high and low density) areas. For precipitation, we decided to use ground measured data. While precipitation data from the Department of Hydrology and Meteorology (DHM) was mainly collected in the centre of the Valley, citizen scientists also measured precipitation at the edges of the Valley. We combined the DHM and citizen science data and interpolated this combined data over the Valley in QGIS using Inverse Distance Weighting (IDW) interpolation. This resulted in a spatial precipitation map covering the whole Valley.

We visualized the relation between the characteristics and specific discharge using scatter plots. We fitted regression lines through the obtained scatter plots to search for relations. Both linear and polynomial relations were tested and the best fit was determined based on the computed coefficient of determination. (R^2). Furthermore, the Pearson’s correlation coefficient (r) between the characteristics and the specific baseflow was computed. A correlation coefficient of 1 indicates a perfect positive correlation, whereas a coefficient of zero shows none. Unfortunately, the relations we found using this technique were not able to compute any sensible data. Looking at the scatter plots in Figure 3, it can be observed that outliers seemed to have a large impact on R^2 and r.

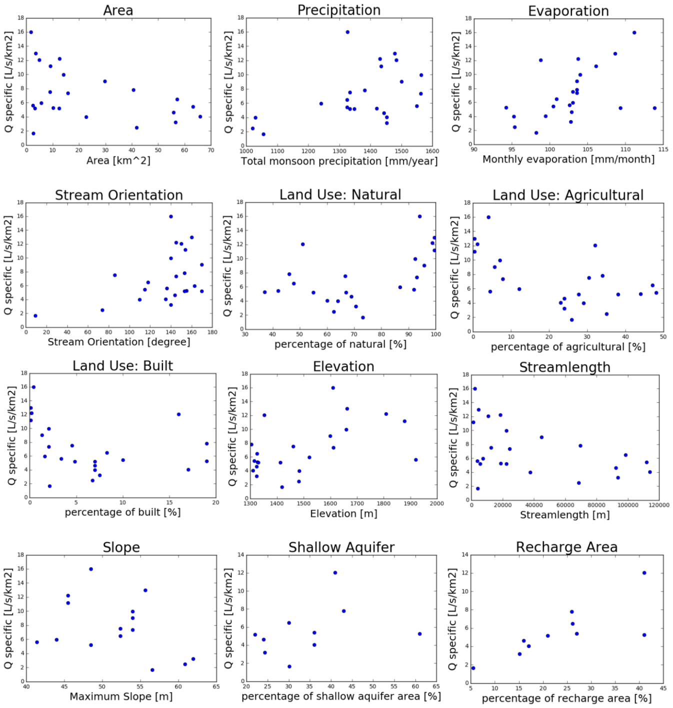

Fig. 3. Scatter plots between characteristics and specific baseflow.

Due to the effect of the outliers, we narrowed the characteristics down to the three most important based on our physical understanding of the relationships between specific baseflow and on the largest correlation coefficients.

Because of the cross-correlation between precipitation and elevation, it was chosen to only include precipitation, knowing that this phenomenon would certainly be of influence. Other cross-correlations between characteristics were also left out of the new model, hoping this would prevent unnecessary noise. Furthermore, we realised that areas with a high percentage of natural and agricultural land have on average a higher infiltration capacity than built areas and thus will contribute more to the baseflow. Additionally, we expected that, due to the strong south to north monsoonal air movement [3], the orientation of the stream and its impact on orographic precipitation would be of importance too.

Ultimately, we decided to train both the Regression Model and the Black Box, by only using precipitation, land-use (natural and agricultural), and stream orientation.

C. *Three methods to approach baseflow*

To predict the baseflow for unknown watersheds we used three different methods: Spatial Analysis, Regression Model,

TABLE I
CHARACTERISTICS AND CORRESPONDING CORRELATION

Characteristic	Corr. Coefficient
Area	-0.43
Precipitation	0.48
Evaporation	0.48
Stream orientation	0.46
Land-use natural	0.50
Land-use agricultural	-0.56
Land-use built	-0.31
Basepoint elevation	0.49
Stream length	-0.44
Slope	-0.26
Shallow aquifer	0.35
Recharge areas	0.78

and Black Box. Below, the methods and the input data that are used are explained and any assumptions made for each method are clarified.

1) Spatial Analysis: Spatial Analysis is the simplest method that makes use of spatial interpolation. Interpolation is the process of using geographic point data to compute values at unknown locations. There are several interpolation methods available. We decided to use Inverse Distance Weighted (IDW) interpolation. IDW is a commonly used method which is also available in QGIS. We used IDW interpolation of specific baseflow in combination with land-use, since we assume that only natural and agricultural land will contribute to the baseflow. Firstly, we uploaded the known specific baseflow data in QGIS and applied IDW interpolation over the Valley area. For the distance coefficient we selected p=3 (indicating how quickly the weight decreases with distance), since this resulted in the most smooth projection. Next, we created a mask layer in QGIS consisting of zeros for built and ones for agricultural and natural areas. By multiplying these two layers we ensured that only natural and agricultural land-uses generated baseflow. This resulted in our final Spatial Analysis map which shows no specific baseflow value for built areas and, depending on the IDW interpolation, a varying specific baseflow value for the natural and agricultural areas (see Figure 4). Spatial Analysis resulted in a specific baseflow value for every single pixel [L/s/km^2]. It was then possible to calculate each pixel's contribution to the baseflow by multiplying the pixel's area by the interpolated specific discharge. Finally, the actual baseflow for each watershed could be computed by summing the baseflow values of all pixels that belonged to each watershed.

2) Regression Model: Regression analysis is a statistical method to determine the influence of one or more independent variables (characteristics) on a dependent variable (specific baseflow). We decided to build two types of regression models: a manual regression model based on least squares method (Regression Model) and a machine learning based regression model (Black Box).

A simple and understandable way to build a regression model is by using the Least Square Method (LSM). LSM is a reliable method to identify the impact of certain independent variables on the topic of interest. Depending on the number of independent variables that are taken into account, the Regression Model determines estimator coefficients ($\hat{x}$). The combination of the estimators and the characteristics results in an equation which can be used to determine the specific baseflow for all unknown watersheds.

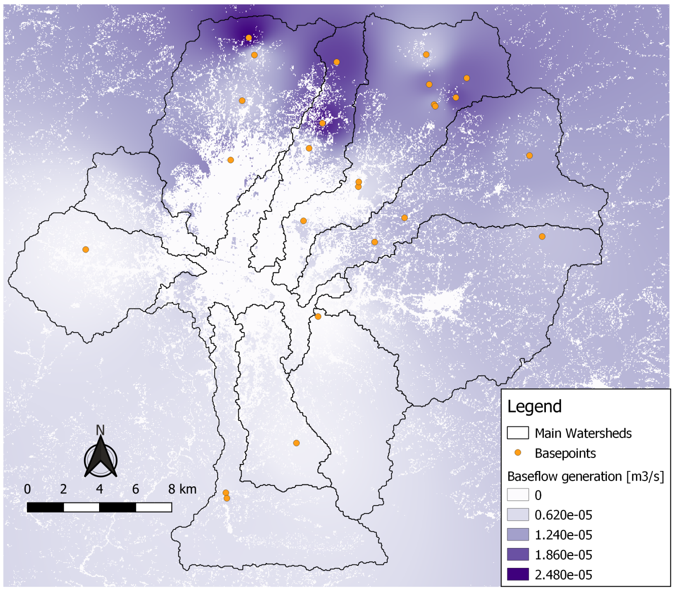

Fig. 4. Map showing the specific baseflow generating areas based on the Spatial Analysis

First, we used the 25 known watersheds to obtain the final equation, since the specific baseflow for those watersheds is known. We created a table in which the characteristics of interest were stored for the known watersheds, called matrix A. The corresponding measured specific baseflow values were stored in a $25x1$ column vector y. The linear relation between matrix A and y is given by Equation 1 where $\hat{x}$ is the estimator.

$$y = A\hat{x} \tag{1}$$

Every characteristic has its own estimator. The estimator can be determined by the least square equation 2.

$$\hat{x} = (A^TWA)^{-1}A^TWy \tag{2}$$

In this equation W is the weight matrix. We assumed all measurements to have the same weight so W is the identity matrix. A^T is the transpose of matrix A. To determine the estimated specific baseflow of a watershed (y$_{est}$), the estimator per characteristic is multiplied with the value for the characteristic that belongs to that watershed. The number of characteristics is given by n (see Equation 3).

$$y_{est} = \hat{x_0} + \hat{x_1} characteristic_1 + ... + \hat{x_n} characteristic_n \tag{3}$$

While optimizing the Regression Model, we first analysed different combinations and relations of characteristics, as defined in Section III-B, to find the combination that led to the best result. First, we performed the 'leave one out' method. 'Leave one out' method means that we trained the model with 24 watersheds as input while leaving one watershed out. A least square equation was used to estimate the specific baseflow of the missing watershed, based on a certain combination of characteristics. We compared the estimated value ($y_{est,i}$) with the known specific baseflow (y_i) resulting in an Absolute Percentage Error (APE). This has been repeated for all 25 watersheds. Figure 5 shows the resulting distributions of the APEs for every tested combination. It can be seen that the first four combinations of characteristics showed a large spread in the APEs. The APE ranges of the other methods were quite comparable, however, there was some variability in the median values of the APEs (shown as dark blue horizontal lines in Figure 5).

Ultimately, the selection of the best result out of these combinations is based on the smallest Mean Absolute Percentage Errors (MAPE), see Equation 4.

$$MAPE = \frac{\sum_{i=1}^{25} \frac{|y_i - y_{est,i}|}{y_i} \cdot 100\%}{25} \tag{4}$$

The MAPEs are shown as green triangles in Figure 5. The best performing combination of characteristics turned out to have a MAPE of 33.3% and it includes: precipitation, land-use agriculture, land-use natural and orientation. To make calculations easier, we combined both land-use types in the land-use "green" (see Equation 5). The estimator $\hat{x}$ for each characteristic was determined. This resulted in the final formula:

$$\begin{aligned} Q_{specific} = &\ 399.50 + 2.81 \cdot 10^{-3} \cdot P - 902.59 \cdot Green \\ &+ 508.55 \cdot Green^2 - 2.10 \cdot \cos{(orientation)} \end{aligned} \tag{5}$$

One remark regarding the procedure of the Regression Model is the following: we decided to drop Godawari and Bhalku watersheds from the training data as they generated aa APE of around 900%. This abnormal high percentage might be explained by unknown industrial activities including groundwater extraction and waste water discharge, resulting in unreliable flow data at these locations. Due to time limits we were not able to find the exact origin of the abnormalities. This decision is also implemented for the Black Box.

3) Black Box: The Black Box Model is a supervised machine learning algorithm. For this research, linear regression via Scikit-Learn, a popular machine learning library for Python [14], was used.

Based on our results with the manual Regression Model we expected to use not only linear but also polynomial relations. However, since the Scikit-Learn is only applicable for linear regression, we had to transform the input characteristics to polynomial coefficients to be able to include polynomial relations as well. To account for the polynomial coefficients, we transformed the normal linear model in Equation 6:

$$h(\theta) = \hat{x}_0 + \hat{x}_1 \cdot \theta_1 + ... + \hat{x}_n \cdot \theta_n \tag{6}$$

As in Equation 3, the 'characteristics' are now defined as the θ-values. The θ-values that gave the best results, were determined:

$\theta_1 = precipitation^2$

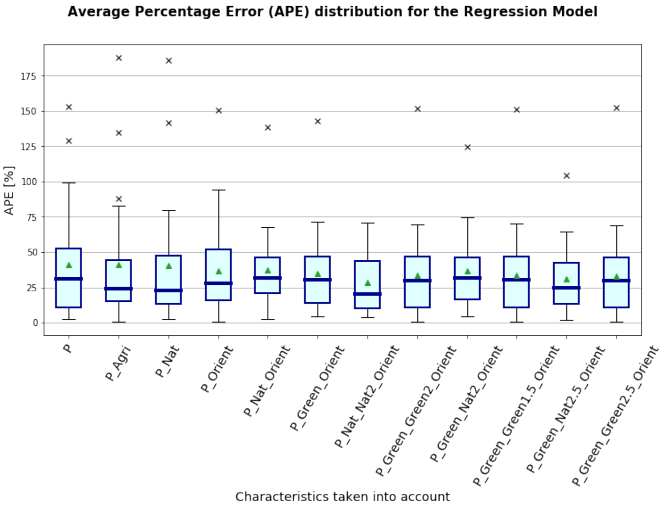

Fig. 5. APE distributions of different characteristic combinations for the Regression Model, including MAPE values (green triangles).

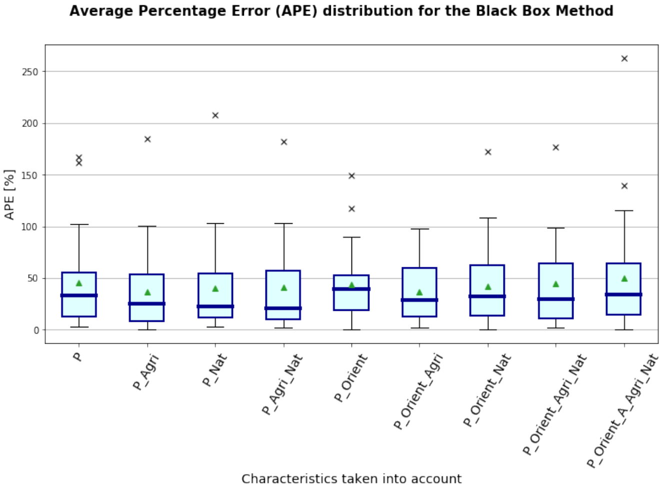

Fig. 6. APE distributions of different characteristic combinations for the Black Box Method, including MAPE values (green triangles).

$\theta_2 = \sqrt{built}$
$\theta_3 = \sqrt{agriculture}$
$\theta_4 = natural^2$
$\theta_5 = \cos{(orientation)}$

Now, we transform Equation 6 to the following polynomial model:

$$\begin{aligned} h(\theta) = &\ \hat{x}_0 + \hat{x}_1\theta_1 + ... + \hat{x}_n\theta_n \\ = &\hat{x}_0 + \hat{x}_1 \cdot precip + \hat{x}_2 \cdot \sqrt{built} + \\ &\hat{x}_3 \cdot \sqrt{agriculture} + \hat{x}_4 \cdot natural^2 + \\ &\hat{x}_5 \cdot \cos(orientation) \end{aligned} \tag{7}$$

This results in:

$$h_\theta(x) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \theta_3 \cdot x_3 + \theta_4 \cdot x_4 + \theta_5 \cdot x_5 \tag{8}$$

Comparable with the Regression Model, Black Box also used a data set to train the model. We used the information about the known watersheds to generate a training matrix

(X_{train}) with the values of different characteristics as the columns of the matrix. The corresponding specific baseflows that were used to train the model were stored in y_{train}. Based on these training matrices, the test matrix (X_{test}) with the characteristics of the unknown watersheds was used to determine the predicted baseflows (y_{pred}). As with the Regression Model, the test phase to determine the best characteristic combination consisted of leaving one watershed out, determining the best coefficients and comparing the estimated specific baseflow with the observed specific baseflow in terms of MAPE.

The comparison of different combinations of characteristics is shown in Figure 6. Based on this figure, the best combination of characteristics was precipitation, orientation, and land-use: agricultural (P_Orient_A_Agri_Nat). This combination resulted in the smallest MAPE of 36.52%. This final combination was used in the next step of this research to determine the specific discharges for the unknown watersheds.

D. *Closing the water balance*

To determine the average groundwater pumping rate ($Q_{\text{pump,net}}$), based on average annual baseflow data, we set up a water balance for the Valley based on research performed by Davids [4]. This water balance is provided in Equation 9:

$$\Delta S = Q_{sw,in} + GW_{pump,net} + P - E - Q_{sw,out} \qquad (9)$$

where $Q_{sw,\,in}$ is the surface water inflow, $Q_{\text{sw,out}}$ is the surface water outflow at Chobar, P is precipitation, E is evaporation, $Q_{\text{pump,net}}$ is the net groundwater pumping and ΔS is the storage change in the unsaturated zone as well as any lakes or reservoirs (which are negligible in the Valley). The water balance domain includes the unsaturated zone from the soil surface down to the water table. For the boundary of our water balance, we used the boundary around the densely built area of Kathmandu (Figure 1.2).

With the available data from S4W-Nepal, DHM, and satellites, we were able to define the baseflow of the surrounding watersheds, which equals $Q_{\text{sw,in}}$ of our water balance. Also the values for $Q_{\text{sw,out}}$, P and E were determined. The precipitation during pre-monsoon is negligible, since care was taken to avoid performing baseflow measurements after precipitation events where surface runoff was still occurring. Finally, we assumed S to be a constant value, so ΔS is zero. For the pre-monsoon months, this is a reasonable assumption, because evaporation is relatively low as well as soil moisture changes. These assumptions result in the final water balance for the Valley, shown in Equation 10:

$$Q_{sw,in} + GW_{pump,net} - E = Q_{sw,out} \qquad (10)$$

IV. Results

Based on the three methods, we determined the baseflow (by converting the specific baseflow) for both the known and unknown watersheds. We then validated and compared the methods by their output for known and unknown watersheds. After summing up the baseflows of smaller watersheds, we were able to determine the total baseflow for the larger watersheds and sum these up for the total surface water inflow to close the water balance.

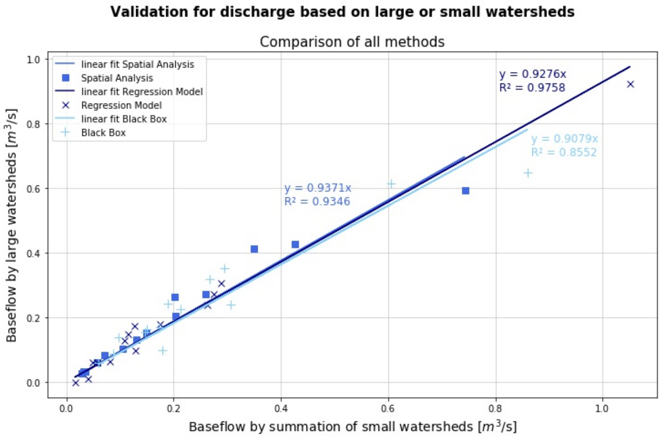

Fig. 7. Validation of the three methods by comparing discharge based on big watersheds is the same as the summation of smaller watersheds

A. *Comparison of the 3 methods*

Using the three methods, we estimated the baseflow $Q[m^3/s]$ for all 67 watersheds (25 known + 42 unknown watersheds). To be able to quantify which method performed best, an extended analysis of the results was performed.

1) Known Q vs estimated Q: For the 25 known watersheds we defined the baseflow (Q [m^3/s]) by the three methods and compared the estimated values with the measured baseflow. In appendix A, we provided a table with all measured and estimated baseflow values for the 25 known watersheds per method. A scatter plot between measured and estimated baseflows shows the performance of the three methods (Figure 7). For a perfect model, we would expect a linear fit where the measured baseflow equals the calculated baseflow ($y = x$). Figure 8 show that all methods are better in estimating the baseflow for watersheds with a smaller baseflow than for those with higher baseflow values. Although the differences between the methods are not too big, the Regression Model seems to give the best approximations. This is because this method approaches the $y = x$-line best. Next, we compared the baseflow distribution for the methods in Figure 9. From the baseflow distribution it can be observed that the measured baseflow values have a smaller spread than the three methods. Spatial Analysis and Black Box give a higher average baseflow estimation than what can be found for the measured baseflow values, whereas the Regression Model gives a lower average.

2) Validation of the methods: In order to validate the methods and see whether they were able to only make good predictions for small watersheds or also for bigger watersheds, we constructed some bigger unknown watersheds that overlap the summation of smaller watersheds. By comparing the estimated baseflow for the bigger watershed with the sum of the estimated baseflow values for the smaller watersheds, we could check the consistence of the three methods. It is important to note that since the bigger watershed is often covering a slightly bigger area than the small watersheds it is not expected to find a perfect 1:1 relation. The results are presented in Figure 7. It can be observed that none of the methods shows clear inconsistency. All methods find sufficient

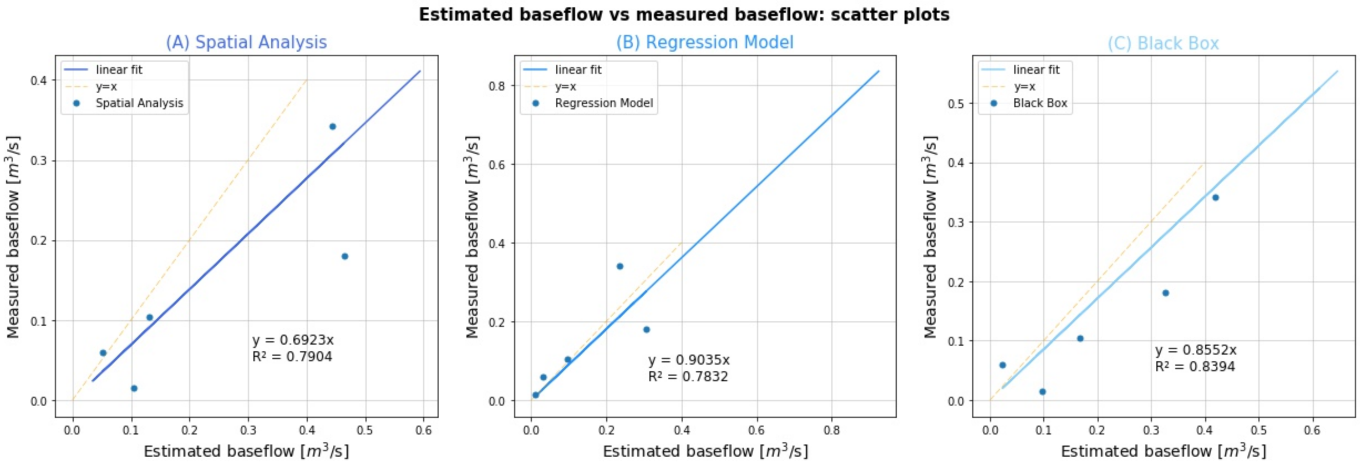

Fig. 8. Scatterplots between estimated and measured baseflows using (A) Spatial Analysis, (B) Regression Model and (C) Black Box for the known watersheds.

comparable baseflow values, so the output by the different methods seems reliable.

3) The range of the output: For the 42 unknown watersheds we defined the baseflow by using the three methods. However, for those watersheds it was not possible to compare the estimated values with measured baseflow data. Figure 10 provides box plots of the estimated baseflow distribution for the unknown watersheds. Numerical values are provided in appendix B. From both the box plots and the numerical values, we observed that the range for the baseflow values are comparable for all methods. The Regression Model shows the highest baseflow estimates, where Spatial Analysis shows the lowest.

4) Final estimated baseflow: Our main interest for the three methods was to find the estimated baseflow for the nine big watersheds in the Valley (see Figure 2). Those watersheds belong to the nine tributaries of the Bagmati River. Their outflow point is located at the edge of the densely urbanised area near Chobar and equals the surface water inflow to our water balance. Based on the three methods, the baseflow for those watersheds was defined and presented in Figure 11. For five out of nine watersheds the baseflow was measured at the outflow point (the purple bar). Even though the three methods are not capable of estimating the exact same baseflow, their values do show orders of magnitude that correspond with the expected values. The numerical values of the estimated baseflow Q [m^3/s] per watershed are presented in table II. It can be seen that the sums for the different methods are quite close to each other, especially for the Spatial Analysis and the Regression Model. The average sum of the estimated annual averaged baseflow flux of the study area is 2.4 m^3/s, which is equal to 0.35 mm/day of runoff.

B. Water balance

To close the water balance as explained by Equation 10, we found the values for the different fluxes using QGIS and SSEBop satalite data.

- ET (pre-monsoon) = 46 mm/month (SSEBop)
- Area = 130298755 m^2

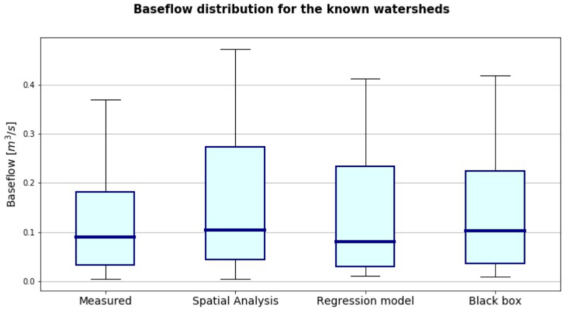

Fig. 9. Baseflow distribution for all methods compared to measured baseflow values for known watersheds

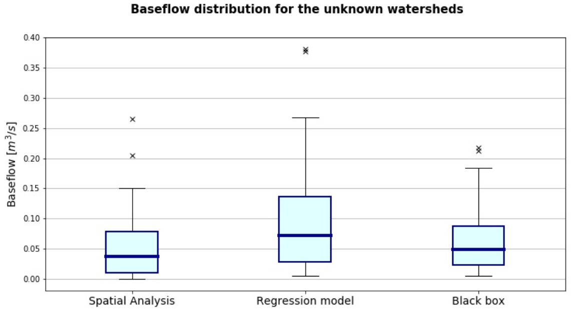

Fig. 10. Baseflow distribution for all methods for unknown watersheds

- ET (pre-monsoon) = 2.3 m^3/s
- $Q_{sw,out}$ (Khokana) = 4 m^3/s
- $Q_{sw,in}$ = 2.4 m^3/s (avg.baseflow 3 methods)

Using Equation 10 and the other fluxes, resulted in a $Q_{net,pump}$ of 3.9 m^3/s for the total water balance area. Taking into account the population growth since 2011, we assumed a population of 3.5 million in the Valley. This resulted in an estimated daily net groundwater pumping of 96 L/person/day during pre-monsoon.

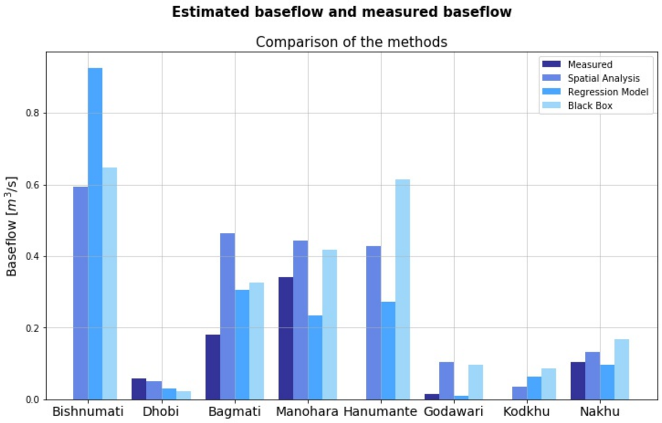

Fig. 11. Baseflow per watershed for three methods

TABLE II
FINAL DISCHARGE [M^3/S] PER WATERSHED

Watershed	Spatial Analysis	Regression Model	Black Box	Average
Balkhu	0.07	0.36	0.27	0.23
Bishnumati	0.59	0.92	0.65	0.72
Dhobi	0.05	0.03	0.02	0.03
Bagmati	0.46	0.31	0.33	0.37
Manohara	0.44	0.23	0.42	0.37
Hanumate	0.43	0.27	0.61	0.44
Godawari	0.10	0.01	0.10	0.07
Kodkhu	0.03	0.06	0.09	0.06
Nakkhu	0.13	0.10	0.17	0.13
Sum:	2.32	2.30	2.65	2.42

V. Conclusion

Here we will be answering the research questions as mentioned in chapter I:

How can monthly streamflow measurements from a subsample of headwater catchments in the Valley be used to determine total pre-monsoon baseflow contributions to the Valleys water balance?

Based on the three methods we evaluated that it is possible to determine pre-monsoon baseflow contributions from a subsample of head water catchments to a certain extent. This was possible by using either Spatial Analysis, a Regression Model, or Black Box. When we look at the results, we can observe that the methods were capable of estimating baseflow values for the known watersheds that were close to the measured baseflow values. This is interesting, since the methods use different techniques and different input data. The scatter plots in Figure 8 show that the methods performed best for watersheds with smaller baseflow values. When we compared the three methods, Regression Model seemed to perform slightly better than the other two. However, the differences in outcome were relatively small. When we validated the methods, we could conclude that the methods find sufficient comparable baseflow values which made the output by the three different methods reliable. Also, for the unknown watersheds, the ranges for the estimated baseflow values were comparable. On average, the Regression Model came up with the highest baseflow values, while Spatial Analysis seemed to give the lowest estimations. Based on the above, it was difficult to point out which method was the best to use in the end. Since the final baseflow estimations were all in the same order of magnitude, we decided to use the average baseflow values of all three methods. The average baseflow contribution of the area outside the water balance by the three methods is 2.4 m^3/s. This average is the best approximation of reality that we could get.

Can these estimates of pre-monsoon baseflow be used in a water balance to estimate net groundwater pumping?

As presented in chapter IV-B, we were able to make an assumption for the net groundwater pumping in the Valley. This calculation was done using the following assumptions:

- There is no precipitation during the pre-monsoon period
- The change in storage (ΔS) was assumed to be a constant
- Recharge to the shallow aquifer was assumed to be zero

Based on a population of 3.5 million people in the densely built areas surrounding Kathmandu, Bhaktapur, and Lalitpur, we found a net ground water extraction of 96 L/person/day during pre-monsoon.

VI. Discussion

One of the main constraints we encountered in this research was the lacking amount of pre-monsoon discharge data. S4W-Nepal did a great effort in collecting discharge and precipitation data over the past three years. However, the amount of measurements per location and the spread of the locations throughout the Valley made it difficult to process the data and train our models in a proper way. The known watersheds were most densely located in the north-eastern part of Valley, and only a few were located in the western and south-eastern region. Because the Regression Model and Black Box used similarities in characteristics to predict the contribution in baseflow, they had difficulties in estimating the baseflow for watersheds in other parts of the Valley than where the measurements were taken. Another issue with the baseflow data was the threshold that we used. Since we had too little data, we were forced to lower the threshold from five to two discharge measurements during the pre-monsoon period, which made the baseflow values less reliable. At the same time, we should also realise that the flow measurements were taken manually by using a SonTek FlowTracker. This measurement technique also introduces uncertainties. One should therefore realise that the final measured baseflow values are only an approach of reality and the final result will never be more precise than the input that we used.

Another point of discussion was the research on the characteristics. In the beginning of our research we found some nice relations between characteristics (linear and polynomial) using BLUE and WLS fit. However, due to the limited amount of data points, several fits were strongly influenced by points that also could be assigned as outliers. This made it doubtful to use the relations for the Regression Model and the Black Box. After we added the eight additional baseflow data points and performed the characteristics research again, the scatter plots between the characteristic and the specific baseflow showed even less correlation and the scatter plots became more cloudy. Due to limited time we were not able to add more characteristics, or perform more in depth research on the existing characteristics. It would have been interesting to add more information about the recharge areas, elevation and soil types and to check their relation with the baseflow. Although, the real problem was not the number of characteristics, but the limited data points. This made it difficult to perform the comparison of characteristics as we had in mind. If one would have more data points, this method could give better results.

To obtain the precipitation data per watershed we used the IDW interpolation method in combination with the measured precipitation data from DHM and S4W-Nepal. IDW interpolation is a common used interpolation method which works well. Although, it might had been better to use Kriging instead. Especially Co-Kriging, with elevation as additional input, would have been preferred. But for convenience we decided to keep working with IDW interpolation.

The method of interpolation was also a point of discussion when we look at Spatial Analysis. We only tested the relation between baseflow and natural land-use and between baseflow and the combination of natural and agricultural land-use. On the other hand, if we had chosen to use Co-Kriging, it would have been possible to add a second parameter to the interpolation technique and thereby obtaining better results. Simultaneously, the main advantage of the Spatial Analysis method is that it needs only little input data to give relative good results.

For the Regression Model and Black Box we compared several characteristics to find the optimal combination that gave the best result. This choice was made based on the combination of characteristics that gave the lowest MAPE value. If, instead, we had chosen to base our decision on the boxplots, another combination of characteristics had maybe been chosen to be the best.

For the Regression Model and the Black Box we decided to remove two known watersheds: Godawari river (Balkot02) and Balkhu river (BK02)) from the training data since they resulted in MAPE values of around 900%. It is still difficult to explain why both models were not capable of modeling the behaviour of those two watersheds. One reason could be the location of BK02, which is the only basepoint that is located in the western part of the Valley. Therefore, the Regression Model and Black Box might have difficulties in predicting the behavior of watersheds in that area. This would also reduce the reliability of their estimations of other unknown watersheds in that region. For Balkot02 groundwater recharge might play a role in disturbing the baseflow estimations. We do not know if this might give problems in other regions as well.

It is remarkable that the watersheds BA05, BA06, MH02 and MH03 were overestimated by all three methods. Those watersheds have in common that they cover a relative big area and that they have a high baseflow value. A reason for the mismatch between the calculated baseflow values and the measured baseflow values could be water extraction for human purposes or the presence of groundwater recharge areas within the watershed.

The three methods all have their pros and cons. Spatial Analysis is the quickest and easiest method to use. Also, when little data is available, it is still able to give relative good results. Regression Model and Black Box need sufficient input data to train their algorithms. This was a problem within our research. Still, although the MAPE is still more than 30% for both methods, we can state that the results were surprisingly comparable, also with respect to the output of Spatial Analysis. An advantage of Regression Model over Black Box is the fact that Regression Model is easier to understand, while the machine learning part still act as a Black Box and does not give all the insights we would desire. We think that when we would have access to more datapoints, Regression Model and Black Box might surpass the results of Spatial Analysis, since their model takes more characteristics into account. However, for now, all methods are performing comparably well.

Finally, we have to discuss the assumptions that were taken for constructing the water balance. To estimate the net groundwater pumping a lot of assumptions were made. We have to assume a relative big error in the baseflow estimations which equals the surface water inflow to our water balance. Furthermore, we did not take any connections between the deep aquifer, the shallow aquifer, and the subsurface flow into account. Also, we were not able to include any information about groundwater extractions. More research on these fluxes would make the estimation of net groundwater pumping more reliable and precise.

VII. Recommendation

Based on our experiences we came up with the following recommendations for further research on the baseflow contributions in the Valley:

Firstly, it would be useful to spread the discharge measurements more throughout the Valley to have more training possibilities for the models. As a result of the spread, the corresponding watersheds will have more differentiating characteristic values (e.g. orientation and land-uses). Furthermore, measuring points located near the boundary of built area would immediately provide baseflow that can be used for setting up the water balance.

Moreover, the recurrence of the flow measurements (once per month) is on the low side. We would suggest to see whether it is possible to generate rating curves for different locations based on the data that is already available. The advantage of rating curves is that citizens can measure water level instead of flow, which is much easier to execute. The water levels can then be related to the discharge based on the rating curve that was generated.

Thirdly, more research is required on the influence of characteristics on the baseflow. Especially the influence of soil types on the baseflow generation needs to be investigated in more detail. Moreover, the spatial map of precipitation could be optimised by using Co-Kriging instead of IDW interpolation.

Next, regarding the Regression Model and Black Box, it would be interesting to see how much better the models perform when more baseflow data is used to train the models.

Finally, the net groundwater estimation could be done more precisely when more research on the different fluxes (e.g.

interaction between the aquifers and the subsurface flow and actual groundwater extractions by companies).

VIII. Acknowledgement

We would like to thank Jeff Davids for his continuous advise, feedback and support. It was because of his enthusiasm that we travelled all the way to Nepal and came in contact with the S4W-Nepal team. Also great thanks to Smartphones4Water(S4W) and the team for hosting us during our research period and providing us with all the data from their measurement campaigns. We had a great time and learned a lot about Nepali and Newari culture. Furthermore, we like to thank Thom Bogaard for giving us the opportunity to start with this multidisciplinary project. Also thanks to the Swedish International Development Agency (SIDA). Finally, this project had not been possible without the financial support of the Delft Deltas, Infrastructures and Mobility Initiative (DIMI).

References

[1] C. B. of Statistics, “National population and housing census 2011 - population projection 2011 2031,” 2011.

[2] ——, “National population and housing census 2011. tech. rep.” 2012.

[3] M. Shrestha, “Interannual variation of summer monsoon rainfall over nepal and its relation to southern oscillation index,” *Meteorology and Atmospheric Physics*, vol. 75, no. 1-2, pp. 21–28, 2000.

[4] J. C. Davids, “Mobilizing young researchers, citizen scientists and mobile technology to close water data gaps,” Ph.D. dissertation, Delft University of Technology, 6 2019.

[5] B. R. Thapa, H. Ishidaira, M. Gusyev, V. P. Pandey, P. Udmale, M. Hayashi, and N. M. Shakya, “Implications of the melamchi water supply project for the kathmandu valley groundwater system,” *Water Policy*, 2019.

[6] B. R. Thapa, H. Ishidaira, V. P. Pandey, and N. M. Shakya, “Impact assessment of gorkha earthquake 2015 on portable water supply in kathmandu valley: Preliminary analysis,” *B1 ()*, vol. 72, no. 4, pp. I_61–I_66, 2016.

[7] B. R. Thapa, H. Ishidaira, T. H. Bui, and N. M. Shakya, “Evaluation of water resources in mountainous region of kathmandu valley using high resolution satellite precipitation product,” *G ()*, vol. 72, no. 5, pp. I_27–I_33, 2016.

[8] B. Thapa, H. Ishidaira, V. Pandey, T. Bhandari, and N. Shakya, “Evaluation of water security in kathmandu valley before and after water transfer from another basin,” *Water*, vol. 10, no. 2, p. 224, 2018.

[9] P. Udmale, H. Ishidaira, B. Thapa, and N. Shakya, “The status of domestic water demand: supply deficit in the kathmandu valley, nepal,” *Water*, vol. 8, no. 5, p. 196, 2016.

[10] V. P. Pandey, S. K. Chapagain, and F. Kazama, “Evaluation of groundwater environment of kathmandu valley,” *Environmental Earth Sciences*, vol. 60, no. 6, pp. 1329–1342, 2010.

[11] D. Pathak, A. Hiratsuka, and Y. Yamashiki, “Influence of anthropogenic activities and seasonal variation on groundwater quality of kathmandu valley using multivariate statistical analysis,” in *Proceedings of the Symposium on Water Quality: Current Trends and Expected Climate Change Impacts*, 2011.

[12] R. Cresswell, J. Bauld, G. Jacobson, M. Khadka, M. Jha, M. Shrestha, and S. Regmi, “A first estimate of ground water ages for the deep aquifer of the kathmandu basin, nepal, using the radioisotope chlorine-36,” *Ground water*, vol. 39, pp. 449–57, 05 2001.

[13] S. S. Vishnu P. Pandey and F. Kazama, “A gis-based methodology to delineate potential areas for groundwater development: a case study from kathmandu valley, nepal,” *Applied water science*, vol. 3, pp. 435–465, 03 2013.

[14] N. Singh Chauhan. A beginners guide to linear regression in python with scikit learn.

Appendix A
Output of three models for known watersheds

		Measured		Spatial Analysis		Regression Model		Black Box	
Layer	Area [m2]	Q [m3/s]	Qspecific [L/s/km2]	Q [m3/s]	Qspecific [L/s/km2	Q [m3/s]	Qspecific [L/s/km2	Q [m3/s]	Qspecific [L/s/km2]
BA01	2485598	0.0139	5.5801	0.0165	2397096	0.0200	8.0508	0.0231	9.2826
NA01	9211998	0.1029	11.1724	0.0921	9197487	0.1038	11.2694	0.1268	13.7700
BA02	14174337	0.1408	9.9356	0.1139	13869999	0.1323	9.3327	0.1376	9.7074
NA02	12677816	0.1547	12.1993	0.1301	12653199	0.1393	10.9890	0.1387	10.9389
BA03	15697733	0.1155	7.3546	0.1280	15360579	0.1480	9.4261	0.1677	10.6829
BA035	29828548	0.2701	9.0551	0.2730	29400930	0.3074	10.3062	0.2397	8.0362
BA05	56630308	0.1809	3.1940	0.4637	52361946	0.3066	5.4141	0.3257	5.7516
BK02	12068213	0.0131	1.0847	0.0180	10470564	0.0131	1.0894	0.0546	4.5209
BM015	1829572	0.0293	16.0201	0.0269	1796301	0.0189	10.3330	0.0150	8.1881
BM02	3291886	0.0170	5.1764	0.0438	3116529	0.0221	6.7236	0.0313	9.5130
DB01	3548433	0.0461	12.9973	0.0434	3543930	0.0407	11.4783	0.0450	12.6863
Balkot02	46199868	0.0149	0.3227	0.1043	42356808	0.0104	0.2251	0.0974	2.1091
HM01	5683875	0.0339	5.9554	0.0350	5614011	0.0526	9.2551	0.0359	6.3099
KK01	2714046	0.0045	1.6433	0.0052	2649582	0.0112	4.1406	0.0088	3.2486
MH01	8989658	0.0674	7.5008	0.0732	8589087	0.0441	4.9043	0.0279	3.1069
MH02	57176385	0.3696	6.4647	0.4142	52428870	0.2389	4.1788	0.3522	6.1602
NK03	41842246	0.1039	2.4824	0.1310	39009087	0.0976	2.3327	0.1684	4.0244
BM_extra1	40634278	0.3162	7.7816	0.3382	32664996	0.3177	7.8181	0.2245	5.5245
BM_extra2	12399538	0.0642	5.1776	0.1243	11571768	0.0719	5.7975	0.0776	6.2574
DB_extra1	4905792	0.0592	12.0572	0.0510	4105179	0.0302	6.1507	0.0232	4.7391
DB_extra2	10175321	0.0538	5.2824	0.0965	8286408	0.0810	7.9618	0.0716	7.0411
BA06	65951319	0.2673	4.0522	0.4728	54789462	0.4130	6.2616	0.3636	5.5126
BA_extra1	55861624	0.2583	4.6239	0.4574	51656202	0.3183	5.6973	0.3124	5.5918
MH03	63256915	0.3421	5.4081	0.4440	56991870	0.2345	3.7074	0.4184	6.6144
NK_extra1	22747555	0.0902	3.9653	0.0734	21119085	0.0808	3.5507	0.1025	4.5039

APPENDIX B
OUTPUT OF THREE MODELS FOR UNKNOWN WATERSHEDS

		Measured		Spatial Analysis		Regression Model		Black Box	
Layer	Area [m2]	Q [m3/s]	Qspecific [L/s/km2]	Q [m3/s]	Qspecific [L/s/km2]	Q [m3/s]	Qspecific [L/s/km2]	Q [m3/s]	Qspecific [L/s/km2]
BK_n1	3523362	unknown	unknown	0.0046	2836665	0.0147	4.1676	0.0123	3.4848
BK_n2	647386	unknown	unknown	0.0006	489762	0.0087	13.4869	0.0079	12.2310
BK_n3	1209115	unknown	unknown	0.0016	1107288	0.0055	4.5532	0.0111	9.1600
BK_n4	3635717	unknown	unknown	0.0058	3035916	0.0068	1.8814	0.0120	3.2980
BK_n5	4500806	unknown	unknown	0.0075	3574350	0.0274	6.0879	0.0215	4.7865
BK_n6	4500831	unknown	unknown	0.0106	4053465	0.0026	0.5871	0.0223	4.9651
BK_n7	1186669	unknown	unknown	0.0029	836550	0.0223	18.7708	0.0082	6.9221
BK_n8	3377436	unknown	unknown	0.0054	1854099	0.2118	62.7225	0.0331	9.7881
BK_n9	24834811	unknown	unknown	0.0349	20922876	0.0628	2.5286	0.1377	5.5430
BK_n10	38046645	unknown	unknown	0.0609	31369104	0.1477	3.8824	0.2411	6.3379
BM_n1	15129375	unknown	unknown	0.1068	12577149	0.0945	6.2453	0.1568	10.3668
BM_n2	3388731	unknown	unknown	0.0224	2795598	0.0181	5.3290	0.0372	10.9681
BM_n3	2894389	unknown	unknown	0.0199	2404701	0.0134	4.6343	0.0190	6.5748
BM_n4	12792267	unknown	unknown	0.0626	8551062	0.3654	28.5677	0.0906	7.0819
BM_n5	242954	unknown	unknown	0.0001	24336	0.0755	310.5642	0.0020	8.3114
BM_n6	13151481	unknown	unknown	0.0416	10951200	0.0667	5.0739	0.1059	8.0553
BM_n7	22319729	unknown	unknown	0.1511	18001035	0.1745	7.8194	0.2258	10.1182
MH_n1	15468603	unknown	unknown	0.1213	14741532	0.1094	7.0742	0.0876	5.6663
MH_n2	12491051	unknown	unknown	0.0811	11964186	0.0651	5.2118	0.0596	4.7685
MH_n3	7625312	unknown	unknown	0.0576	6817122	0.0358	4.6935	0.0620	8.1334
MH_n4	10490484	unknown	unknown	0.0863	9332856	0.0465	4.4333	0.0776	7.3970
MH_n5	28256016	unknown	unknown	0.2042	26964288	0.1805	6.3869	0.1551	5.4894
HM_n1	22570133	unknown	unknown	0.1304	20738835	0.0854	3.7856	0.1709	7.5707
HM_n2	12322300	unknown	unknown	0.0710	11151972	0.0225	1.8267	0.0959	7.7836
HM_n3	11209622	unknown	unknown	0.0604	10642437	0.0203	1.8148	0.0446	3.9783
HM_n4	46829603	unknown	unknown	0.2649	43051905	0.1285	2.7432	0.3194	6.8203
HM_n5	4389234	unknown	unknown	0.0267	4071717	0.0211	4.8100	0.0473	10.7752
HM_n6	7299342	unknown	unknown	0.0428	6287814	0.0353	4.8394	0.1023	14.0176
HM_n7	14198468	unknown	unknown	0.0834	12514788	0.0626	4.4104	0.1637	11.5303
HM_n8	19636743	unknown	unknown	0.1080	16524144	0.0925	4.7118	0.2112	10.7533
HM_n9	3220507	unknown	unknown	0.0148	2453373	0.0397	12.3175	0.0342	10.6243
HM_n10	10108209	unknown	unknown	0.0408	9436284	0.0129	1.2768	0.0491	4.8593
GW_n1	18591020	unknown	unknown	0.0479	18084690	0.0854	4.5950	0.0552	2.9686
GW_n2	9557474	unknown	unknown	0.0261	8879598	0.0153	1.5979	0.0356	3.7265
GW_n3	8995658	unknown	unknown	0.0175	8009586	0.0040	0.4488	0.0794	8.8303
GW_n4	35534478	unknown	unknown	0.0863	33071103	0.0365	1.0264	0.0999	2.8103
KK_n1	1647418	unknown	unknown	0.0025	1419093	0.0022	1.3609	0.0085	5.1473
KK_n2	9714405	unknown	unknown	0.0208	8870472	0.0028	0.2918	0.0421	4.3353
KK_n3	3838260	unknown	unknown	0.0064	2699775	0.0634	16.5062	0.0281	7.3285
KK_n4	14377150	unknown	unknown	0.0281	13057785	0.0000	0.0024	0.0581	4.0446
NK_n1	14950149	unknown	unknown	0.0437	14012973	0.0428	2.8596	0.0622	4.1583
NK_n2	3725803	unknown	unknown	0.0124	3487653	0.0041	1.0891	0.0099	2.6452