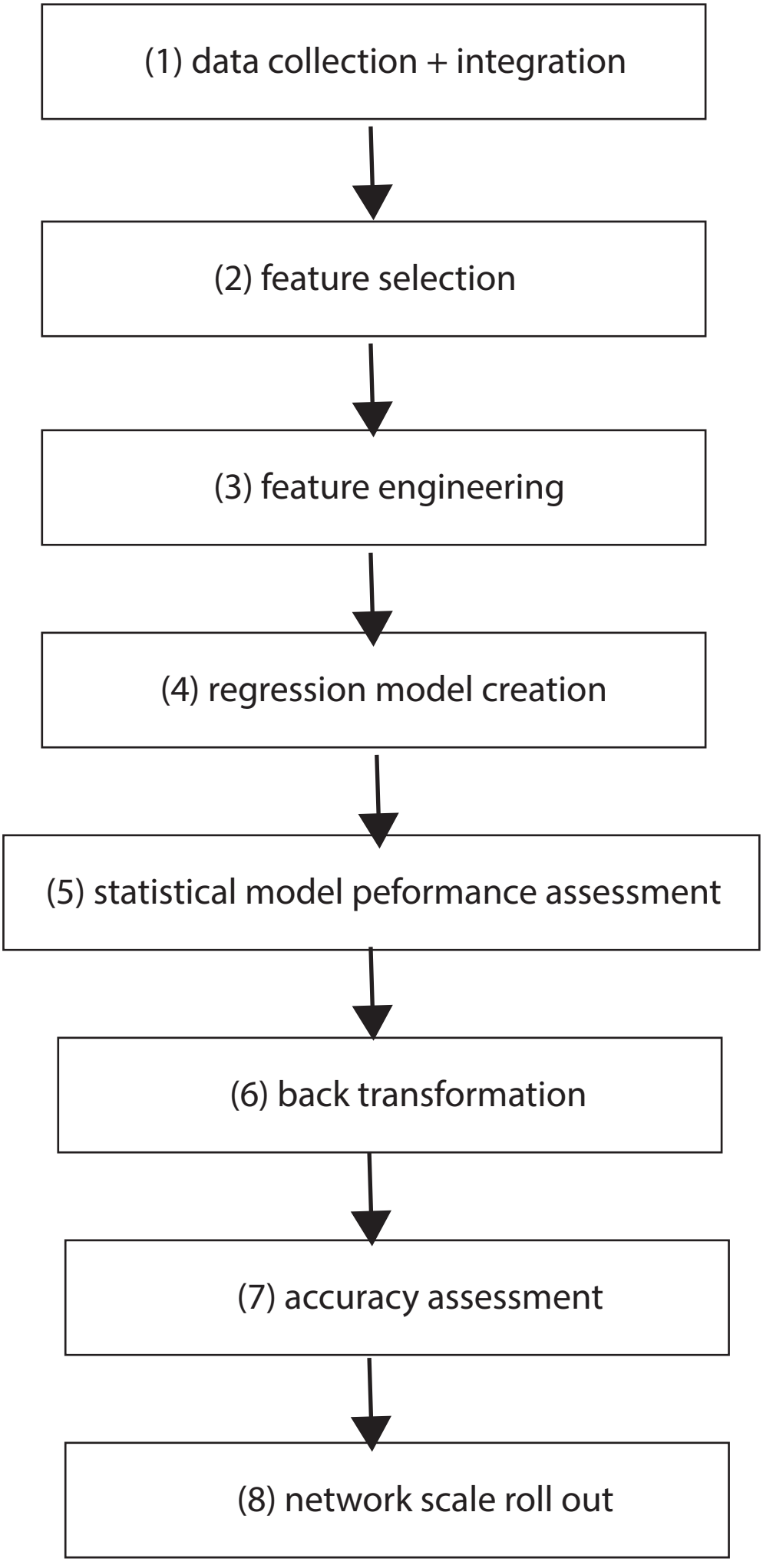
(1) data collection + integration
(2) feature selection
(3) feature engineering
(4) regression model creation
(5) statistical model peformance assessment
(6) back transformation
(7) accuracy assessment
(8) network scale roll out

(1) data collection + integration

Field data (Colin's data) and GIS data (Colin and Hervé' s database)

* I spatially joined the two datasets via SiteID- e.g. SFE_2017_10

(2) feature selection

Phase 1: Segregating the above database by dividing the variables into

(a) dependent variables/ predictions/ response

(b) independent variables/ predictors/ controls

* I used dependent variables as channel cross-section metrics as (i) these cannot be easily measured along the network using GIS and (ii) these GCS metrics are needed for creating archetypes in RB.

* I used independent variables as network scale controls that have been calculated along the entire drainage network

Phase 2: Correlation and covariance analysis on the independent variables. As one of the pre requisite of OLS regression is that the independent variables should not be highly correlated with each other.

* After correlation and covariance analysis of the SFE dataset, I narrowed down my independent variables to 4 variables: contributing catchment area, thalweg slope, confinement and RUSLE (sediment supply)

(3) feature engineering

Phase 1: Data cleaning by visual reconnaissance of the database to identify inconsistencies in the dataset.

* I found that at many places value of confinment was 0 and it is not physically possible as minimun valley width has to be at least 1 km if channel is draining that reach. Therefore, I added value 2 throughout.

* In geomorphology, for network scale analysis, it is common practice to replace slope values < 0.0001 to this value. For channel to flow, there has to be sufficent slope, however, slope extracted from DEMs can result in anaomalous values at relatively flat terrains. Therefore, I replace slope < 0.0001 to this value.

Phase 2: Convert slope from m/m to percent.

Phase 3: Data transformation is needed as one major assumption of OLS regression analysis is that the data should have a gaussian distribution. However, natural world datasets are heavily skewed to the right. Therefore data transformation is needed to get the bell shape of normally distributed dataset.
In river science, it is a commom practice to convert dataset into logarithmic scale. Also, most sclaing relationships studied till date have shown a good relationship using exponential dsitribution. Another transformation that can be used for generating scaling relationships is Box-Cox transformation. Both log and Box-Cox yield decent gaussian distribution in the SFE dataset.

Phase 4: Outlier assessment
It is a common practice to do an outlier assessment before deploying a machine learning model.
* I created a box and rug plot for the width and depth dataset to identify data points lying beyond 3 sigma value and clipped that from the dataset to be processed further.

(4) regression model creation

Phase 1: Perform n * m OLS multiple linear regression runs,
where n is 2: width and depth
m is 4: contributing catchment area (Ac), thalweg slope (s), confinement (cnf) and RUSLE (sediment supply)
* I found that the highest r sq. was obtained when all four of these variables were used as independent variables.

Phase 2: Create 6 bootstrapped regression relationships for predicting width and depth:
width = f (Ac, s, cnf, RUSLE): average, low SD, high SD
depth = f (Ac, s, cnf, RUSLE): average, low SD, high SD

(5) statistical model peformance assessment

R-squared: higher value indicate better model performance i.e. ability to confidently predict
prob of F-statistic: value <0 indicate overall statistically significant relationship between dependent and independent variables
t-statistic: tells the relative importance of variables in the equation and p value of t statistic tells it's significance
Durbin-Watson: measurement of homoscedasticity, or an even distribution of errors throughout our data. Heteroscedasticity would imply an uneven distribution, for example as the data point grows higher the relative error grows higher. Ideal homoscedasticity will lie between 1 and 2.

(6) making predictions and back transformation

Phase 1: Apply the 6 regression equations on the transformed dataset to predict width and depth (average, low SD, high SD).

Phase 2: Back tranform these six series of predictions to obtain predicted width and depth in meter

* average, low SD and high SD predictions are used so as to genrate an envelope of predicted width and depth ranges for each reach

* to back transform log (base 10) transformed data:
back transformed x = 10^(transformed x)
* to back transform Box-Cox transformed data: lambda value generated during original transformation of each series is used

(7) accuracy assessment

deviation of predicted from actual = (predicted-actual)/predicted
* positive values indicate over prediction and
negative values indicate under prediction
lower values indicate more accurate predictions

accuracy of predicted values = 100 - modulus of deviation

(8) network scale roll out

Phase 1: Data cleaning in the GIS attribute table of the drainage network with 200 m segments that are already impregnated with contributing catchment area (Ac), thalweg slope (s), confinement (cnf) and RUSLE (sediment supply)

*I preprocessed the slope and confinment values

Phase 2: Create multiple columns in attribute table for separate columns to transform these four independent variables and predict series of width and depth using the 6 regression equations developed.

Phase 3: Transform the network scale series of the independent variables

Phase 4: Predict normalized network scale width and depth using the 6 regression equations developed.

Phase 5: Back transform the width and depth series to generate predictions in meters