Supplementary Materials: Generalized Linear Models to Identify Key Hydromorphological and Chemical Variables Determining the Occurrence of Macroinvertebrates in the Guayas River Basin (Ecuador)

Minar Naomi Damanik-Ambarita, Gert Everaert, Marie Anne Eurie Forio, Thi Hanh Tien Nguyen, Koen Lock, Peace Liz Sasha Musonge, Natalija Suhareva, Luis Dominguez-Granda, Elina Bennetsen, Pieter Boets and Peter L. M. Goethals

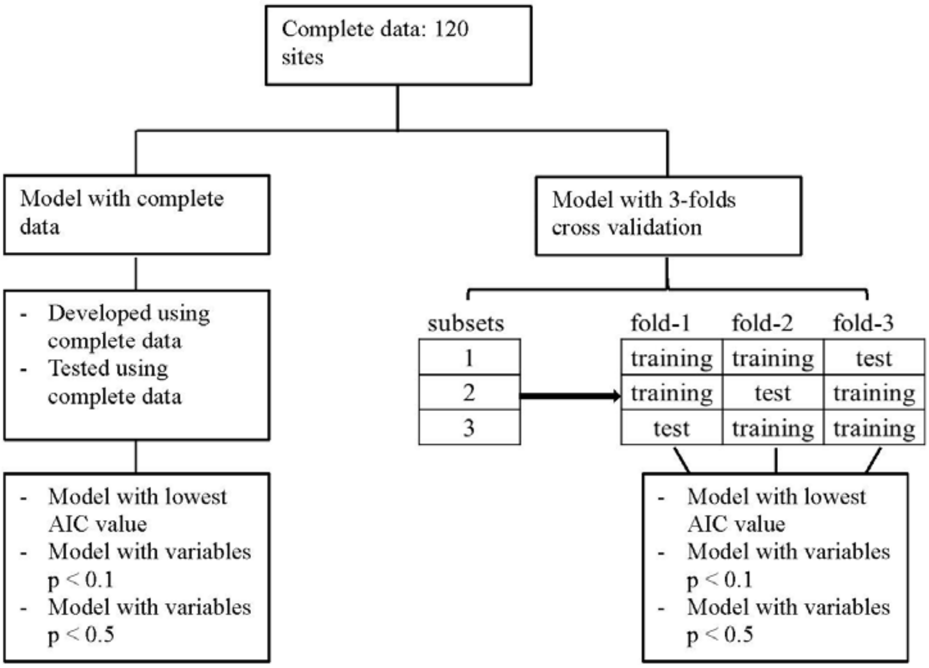


Figure S1. Scheme for model development and criteria for the final models.

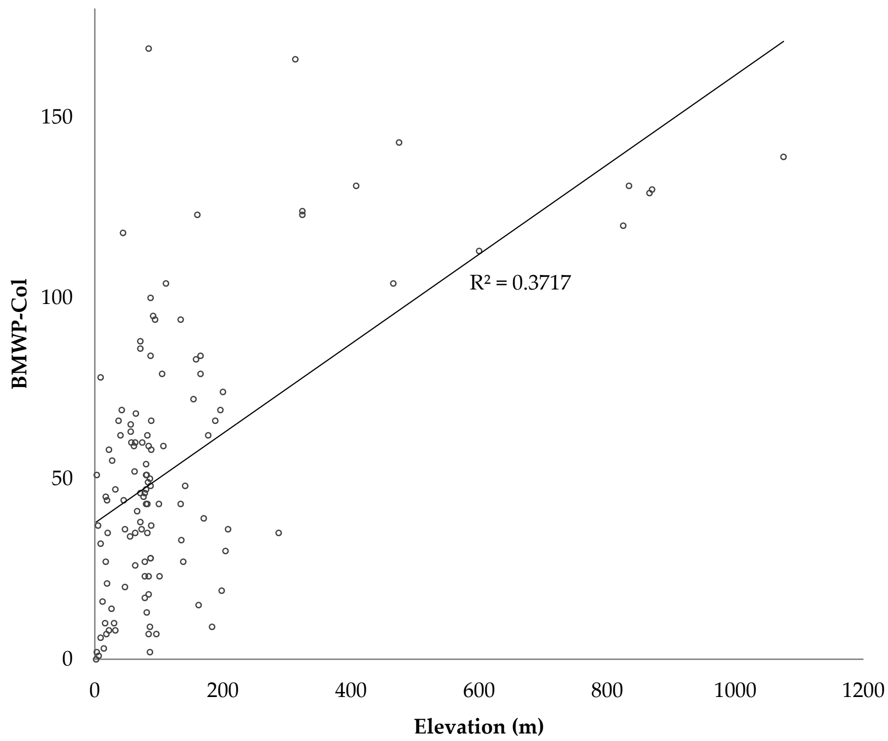


Figure S2. The biotic integrity expressed as the BMWP-Col of sites at different elevations.

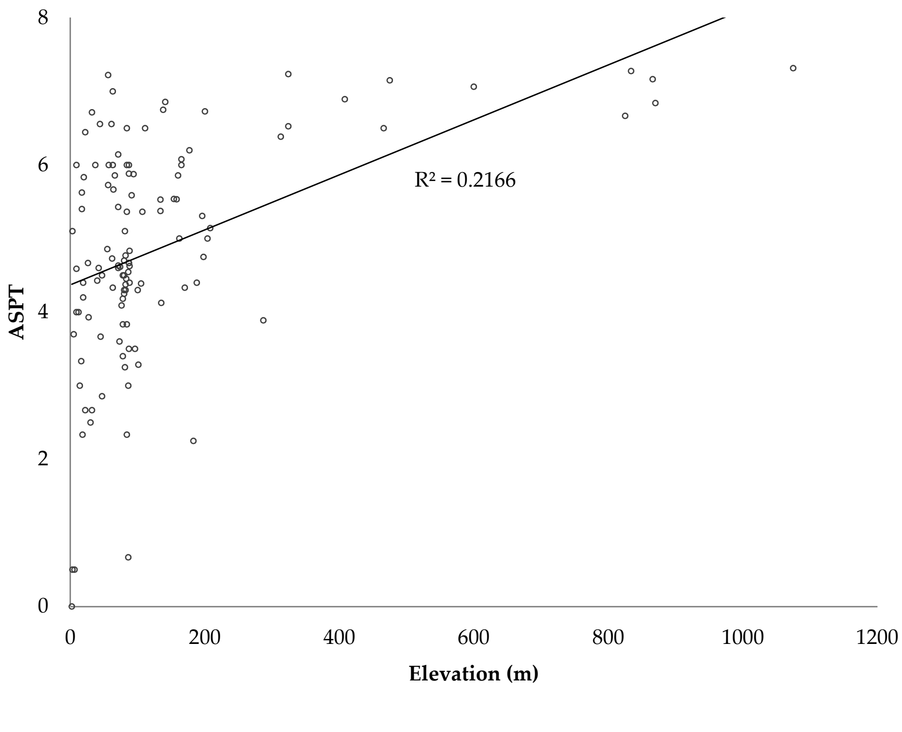


Figure S3. The biotic integrity expressed as the ASPT of sites at different elevations.

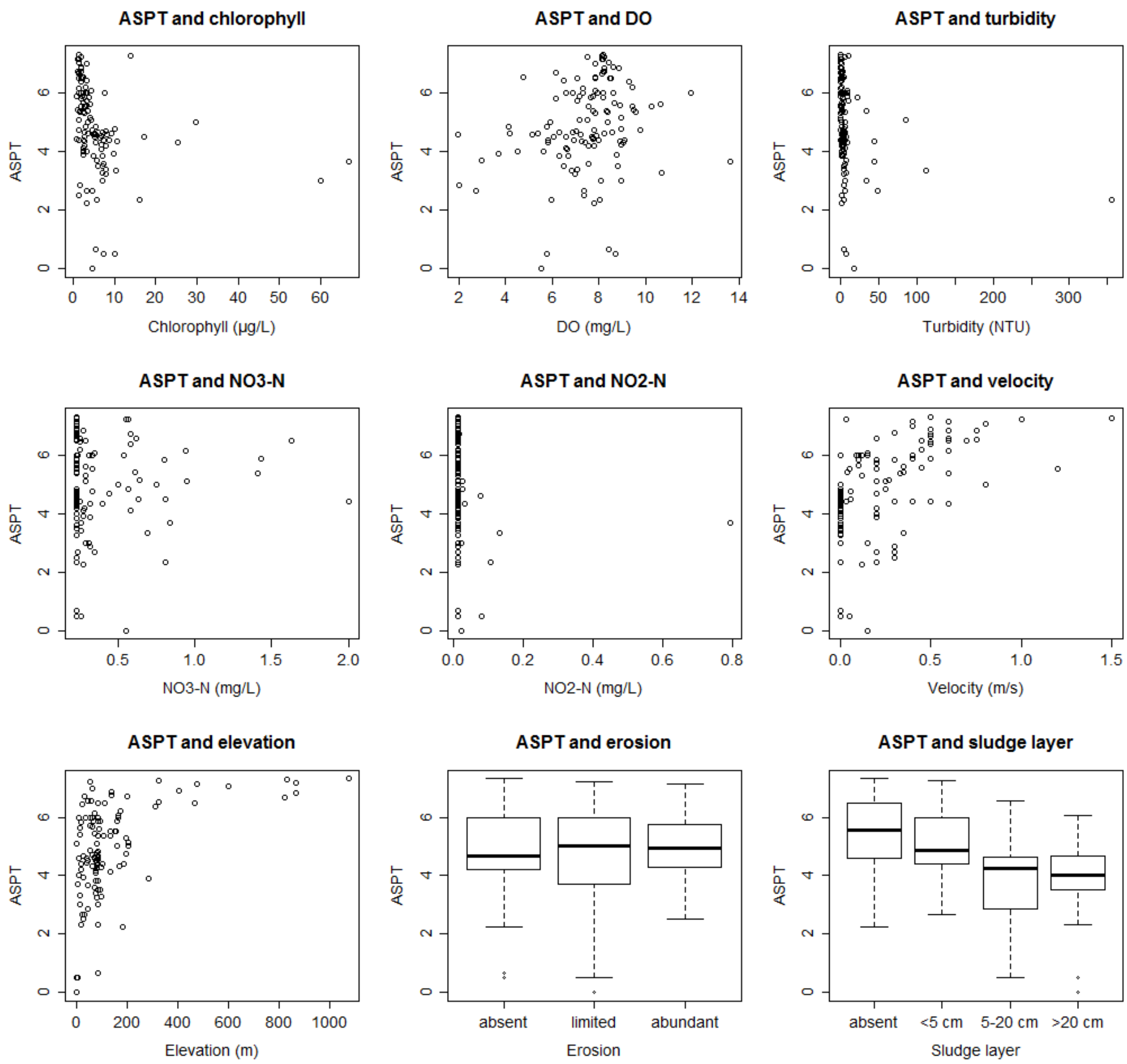
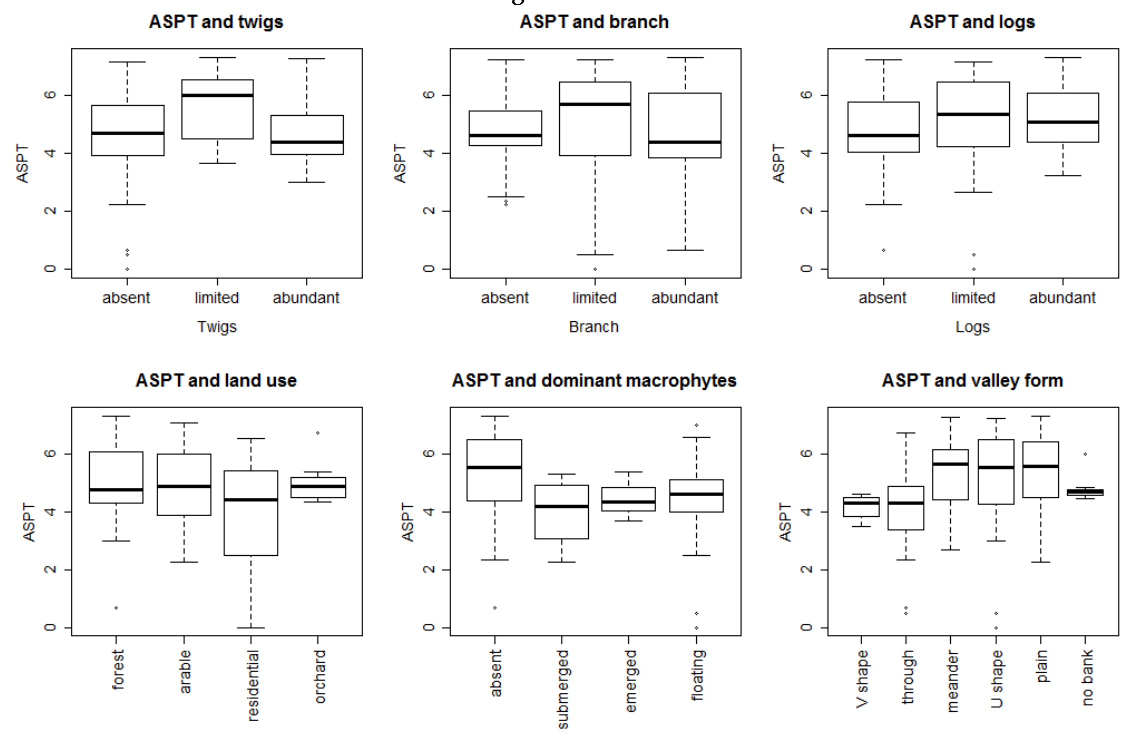


Figure S4. *Cont.*



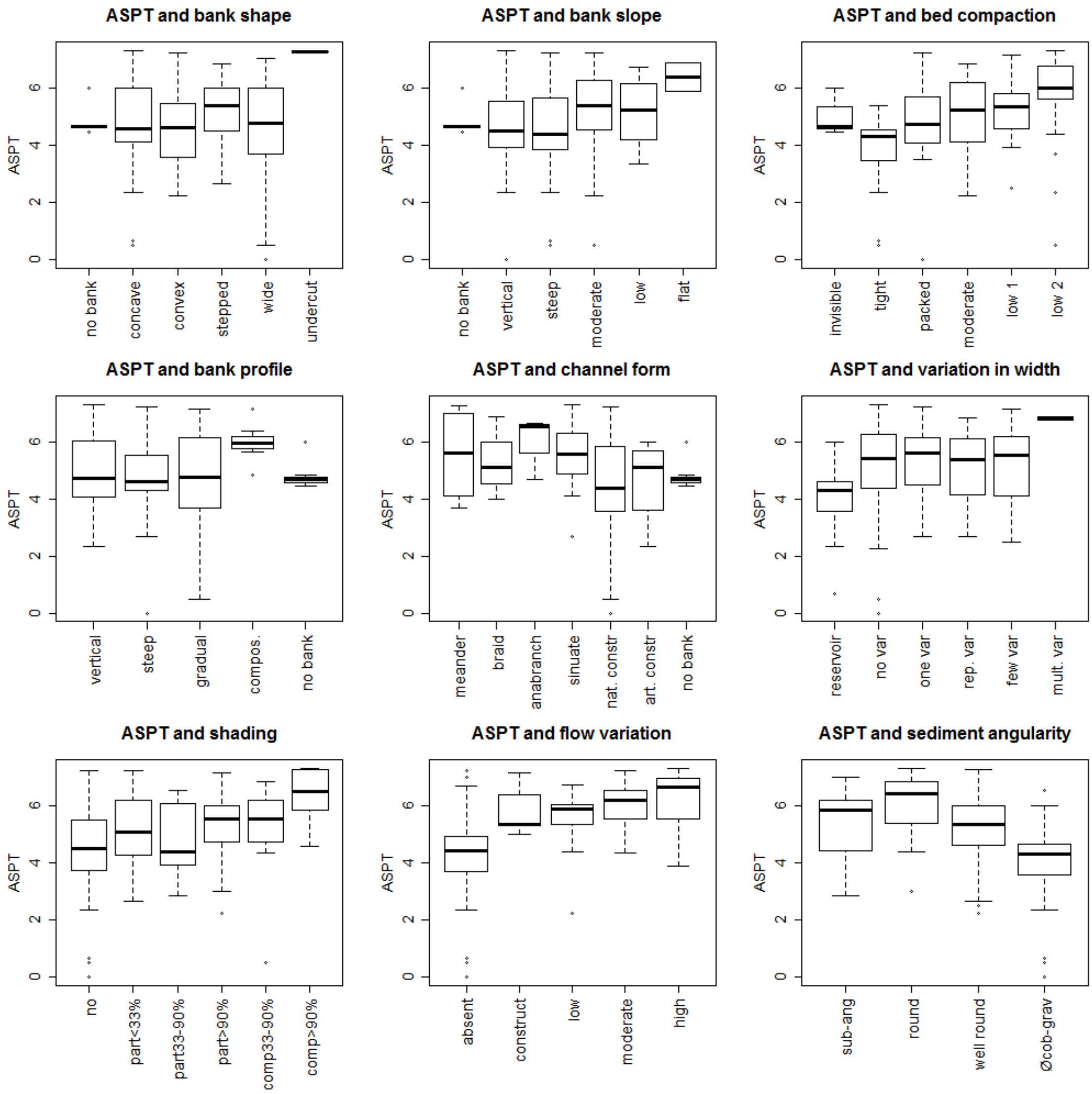


Figure S4. Relationship between environmental variables and the biotic integrity expressed as ASPT, classification of categorical variables are based on Table 2; compos: composite, nat: natural, art: artificial, constr: construction, var: variation, part: partly, comp: completely, ang: angular, cob-grav: cobble-pebble-gravel.

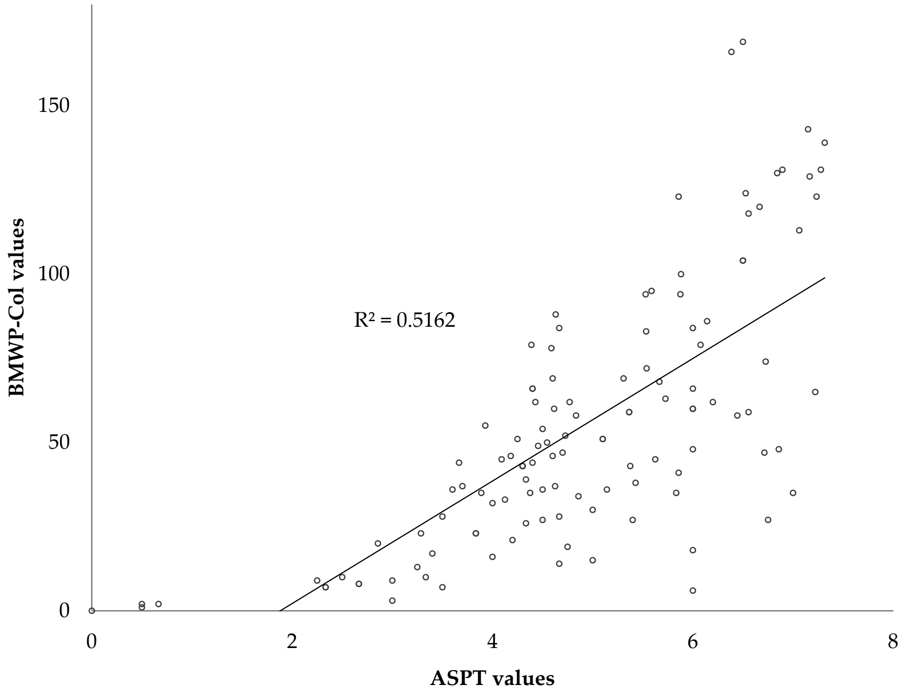


Figure S5. Correlation between BMWP-Col and ASPT and its correlation coefficient.

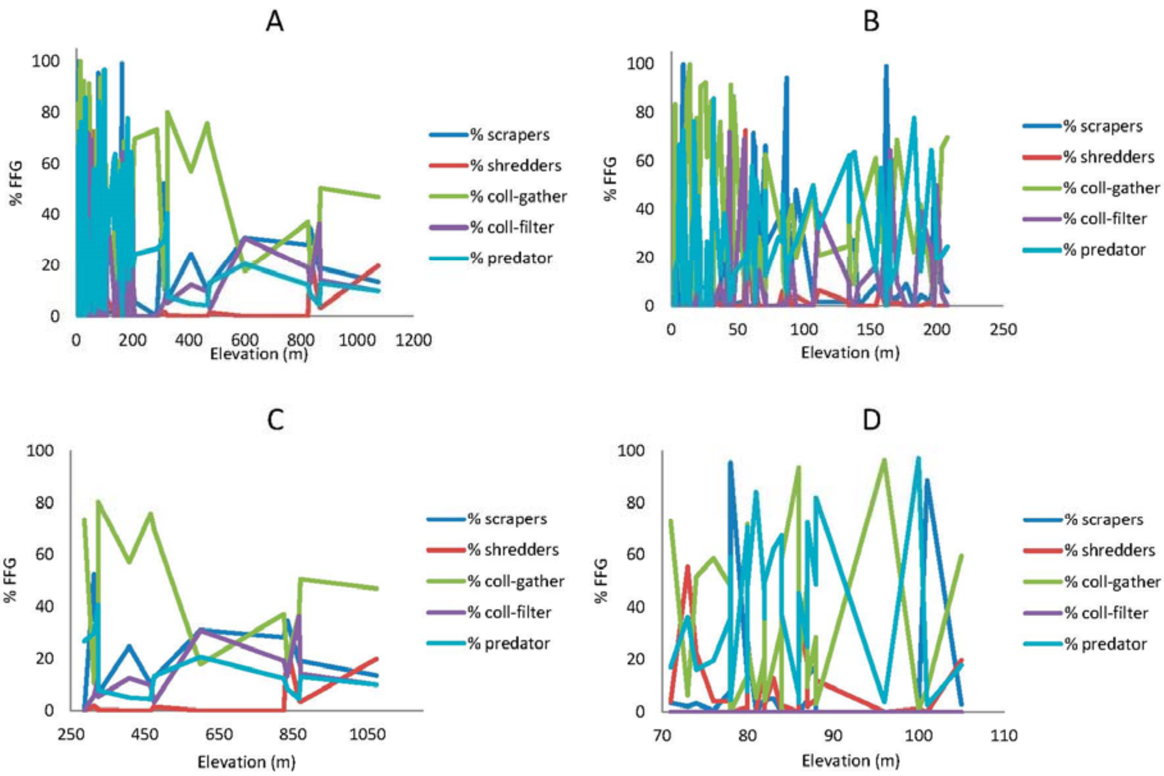


Figure S6. Percentage of functional feeding group (FFG) comprises percentage of scrapers, shredders, collector-gatherer, collector-filterer and predator encountered at the sampling sites: for 120 sampling sites (**A**); for sites located at the elevation lower than 250 m (**B**); for sites located at the elevation higher than 250 m (**C**) and for sites located at the reservoir (**D**).

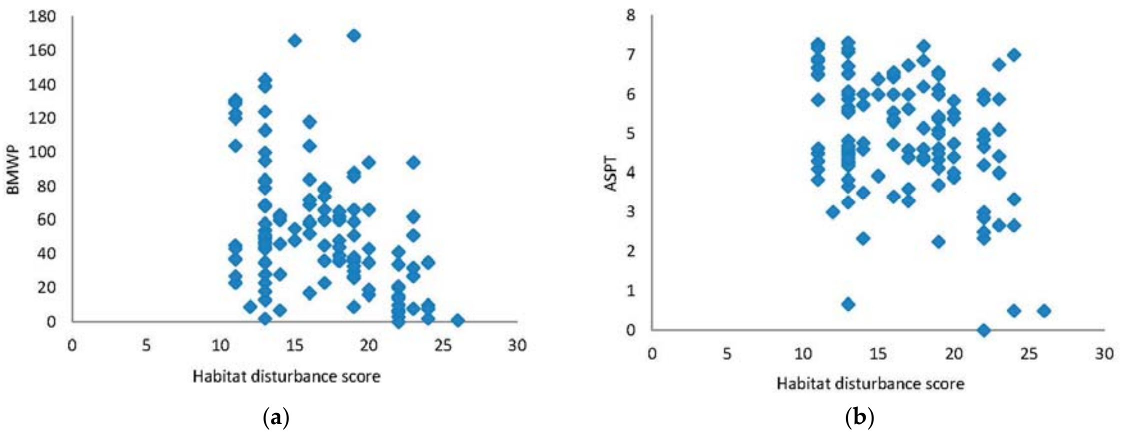


Figure S7. Habitat disturbance score in relation with BMWP-Col (**a**) and ASPT (**b**).

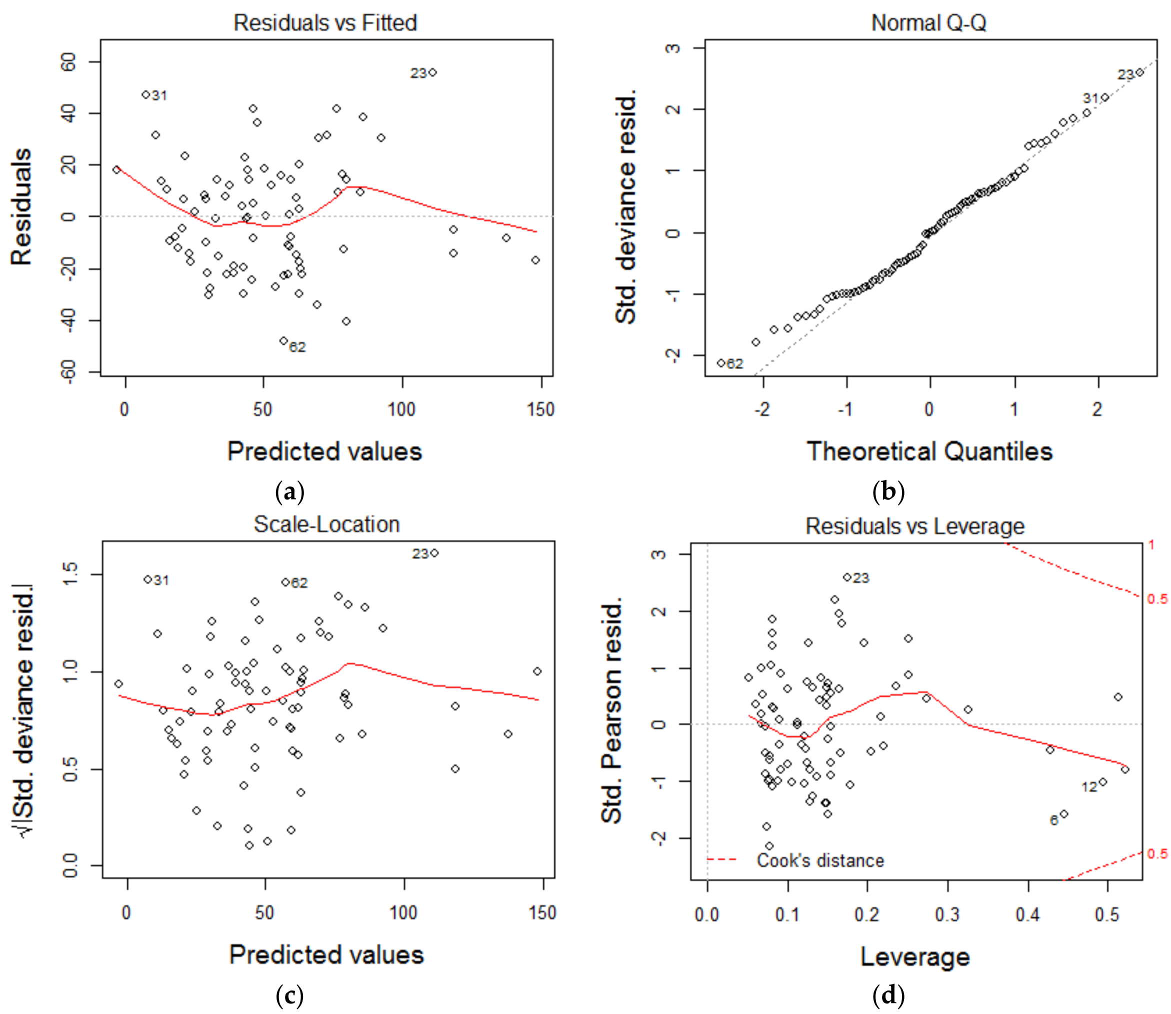


Figure S8. Residuals plots of model based on folds training set 1 + 2 with lowest AIC, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

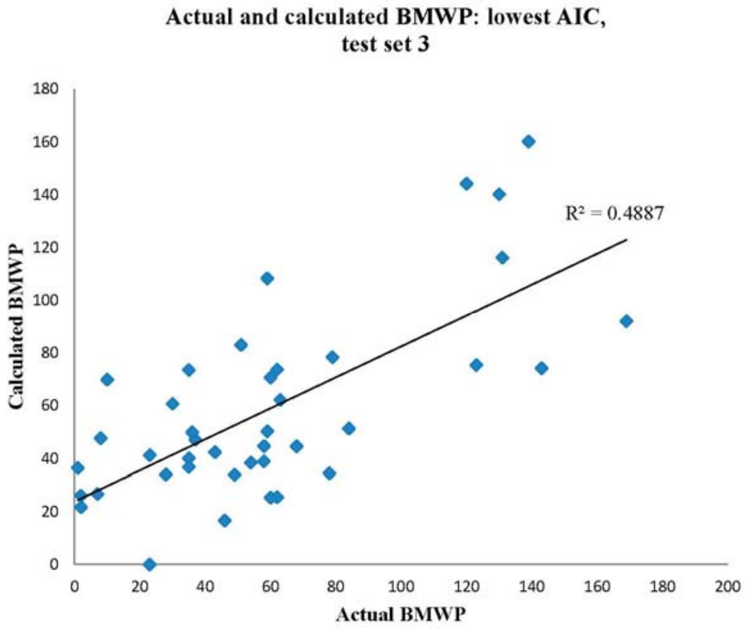


Figure S9. Validation of model based on folds test set 3 with lowest AIC.

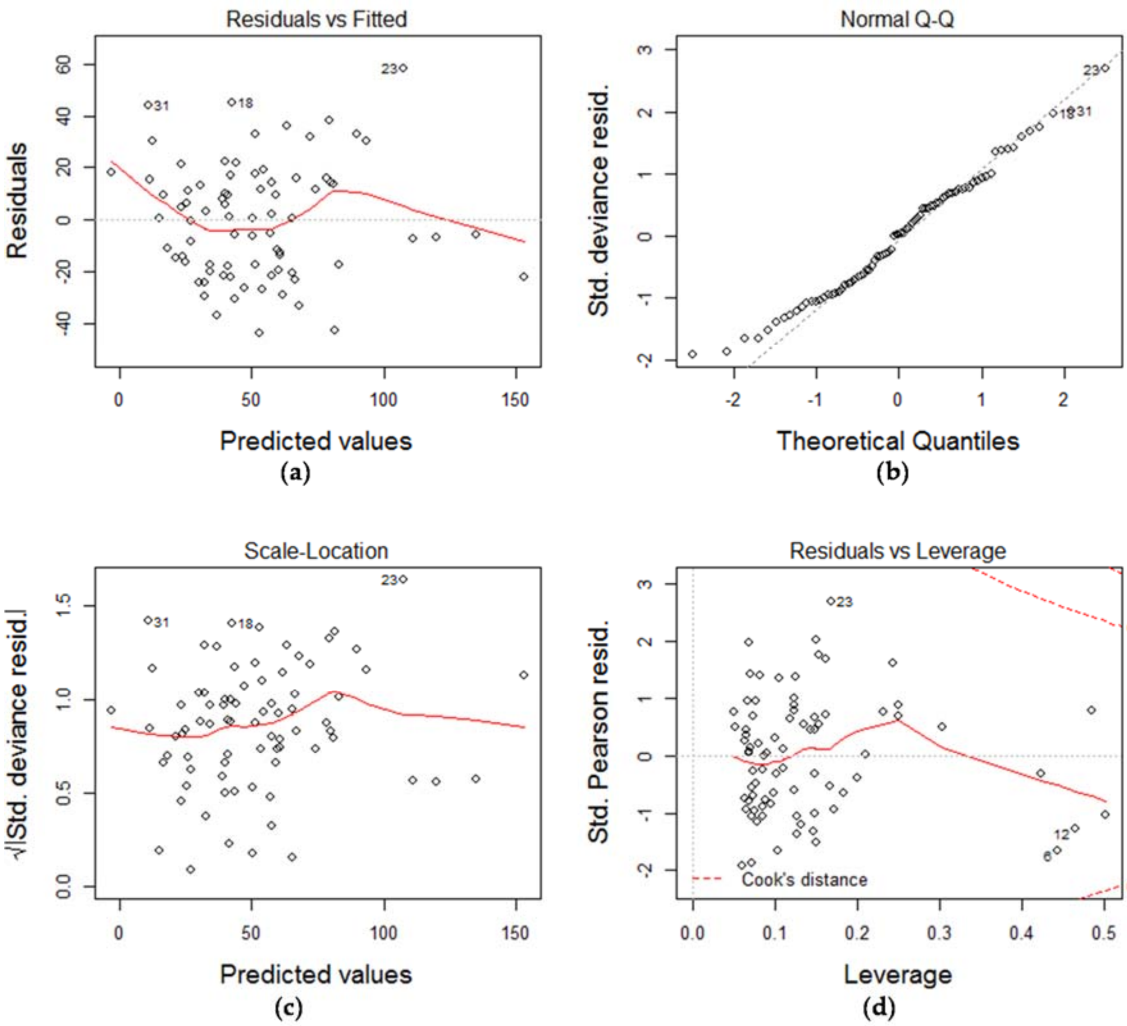


Figure S10. Residuals plots of model based on folds training set 1 + 2 with input variables significant at $p < 0.1$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

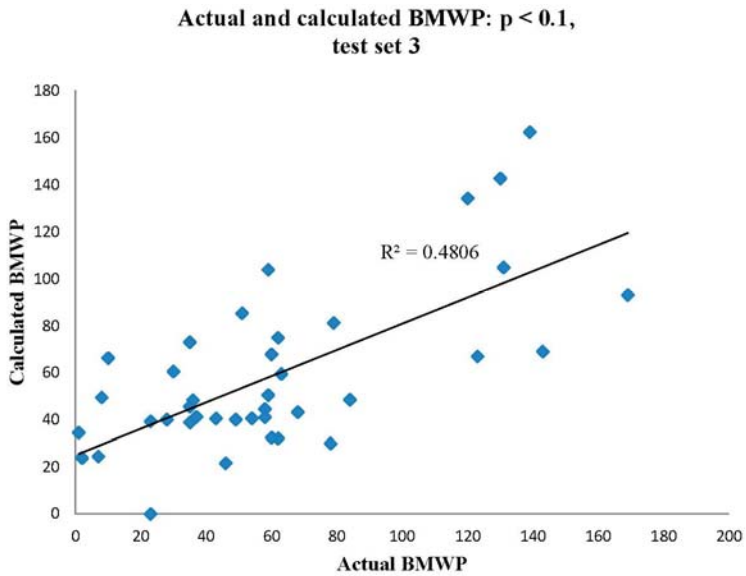


Figure S11. Validation of model based on folds test set 3 with input variables significant at $p < 0.1$.

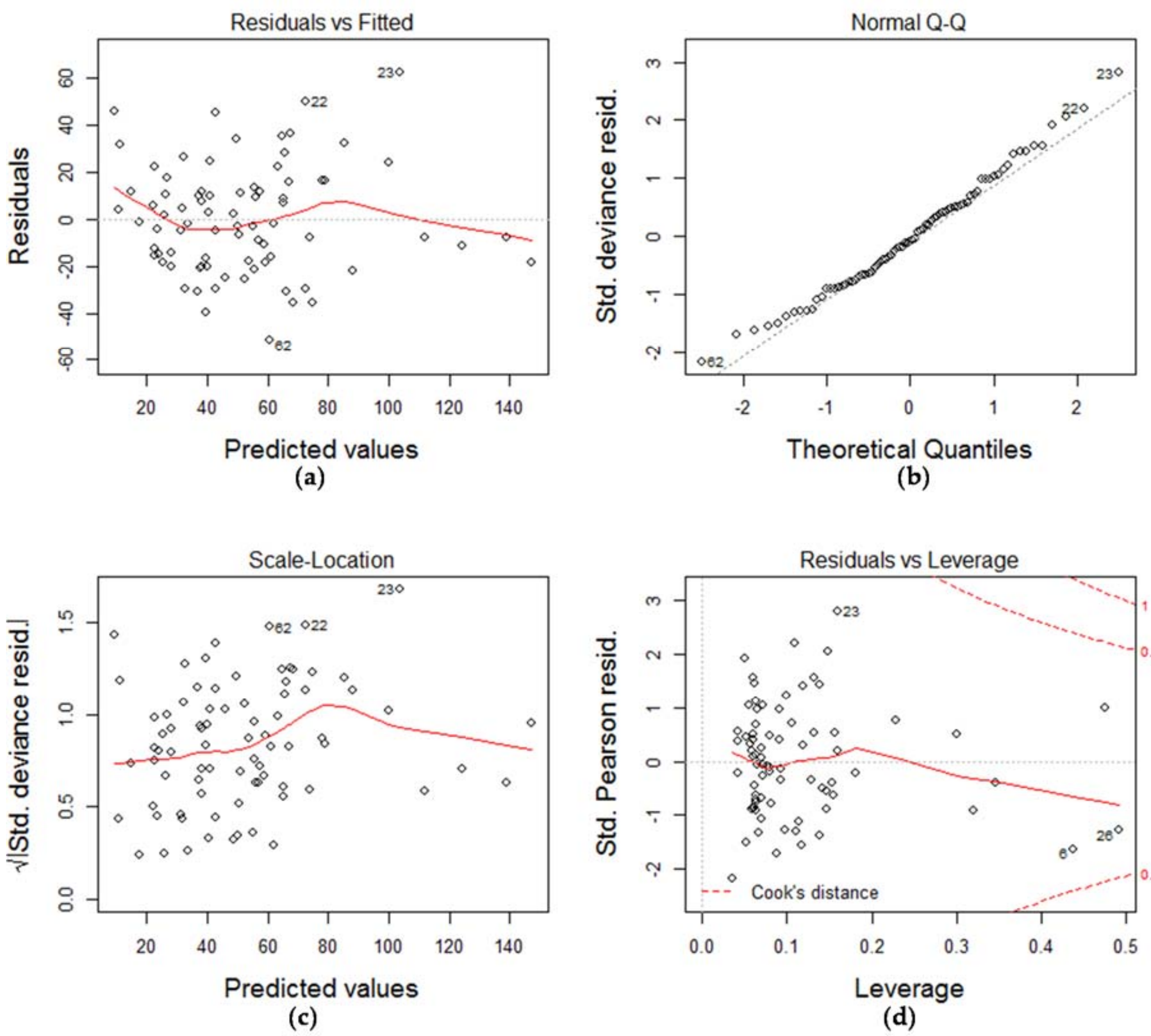


Figure S12. Residuals plots of model based on folds training set 1 + 2 with input variables significant at $p < 0.05$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

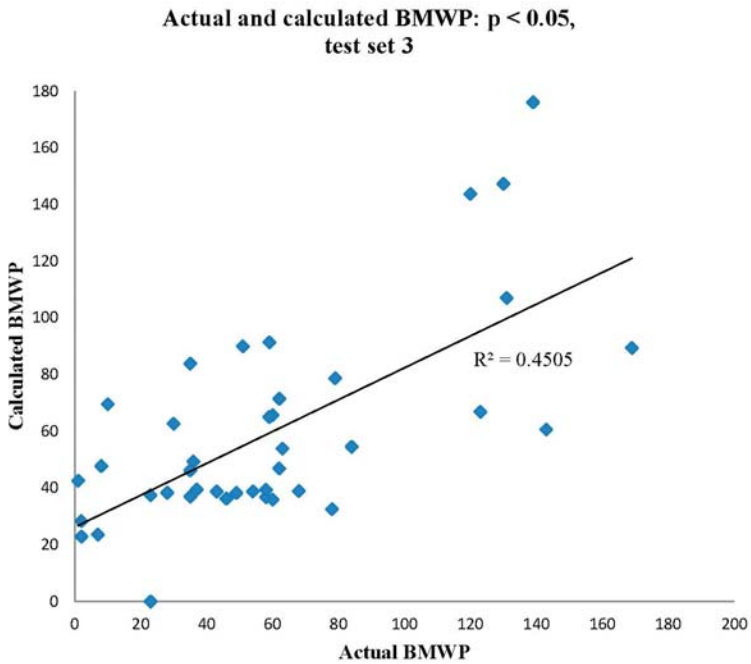


Figure S13. Validation of model based on folds test set 3 with input variables significant at $p < 0.05$.

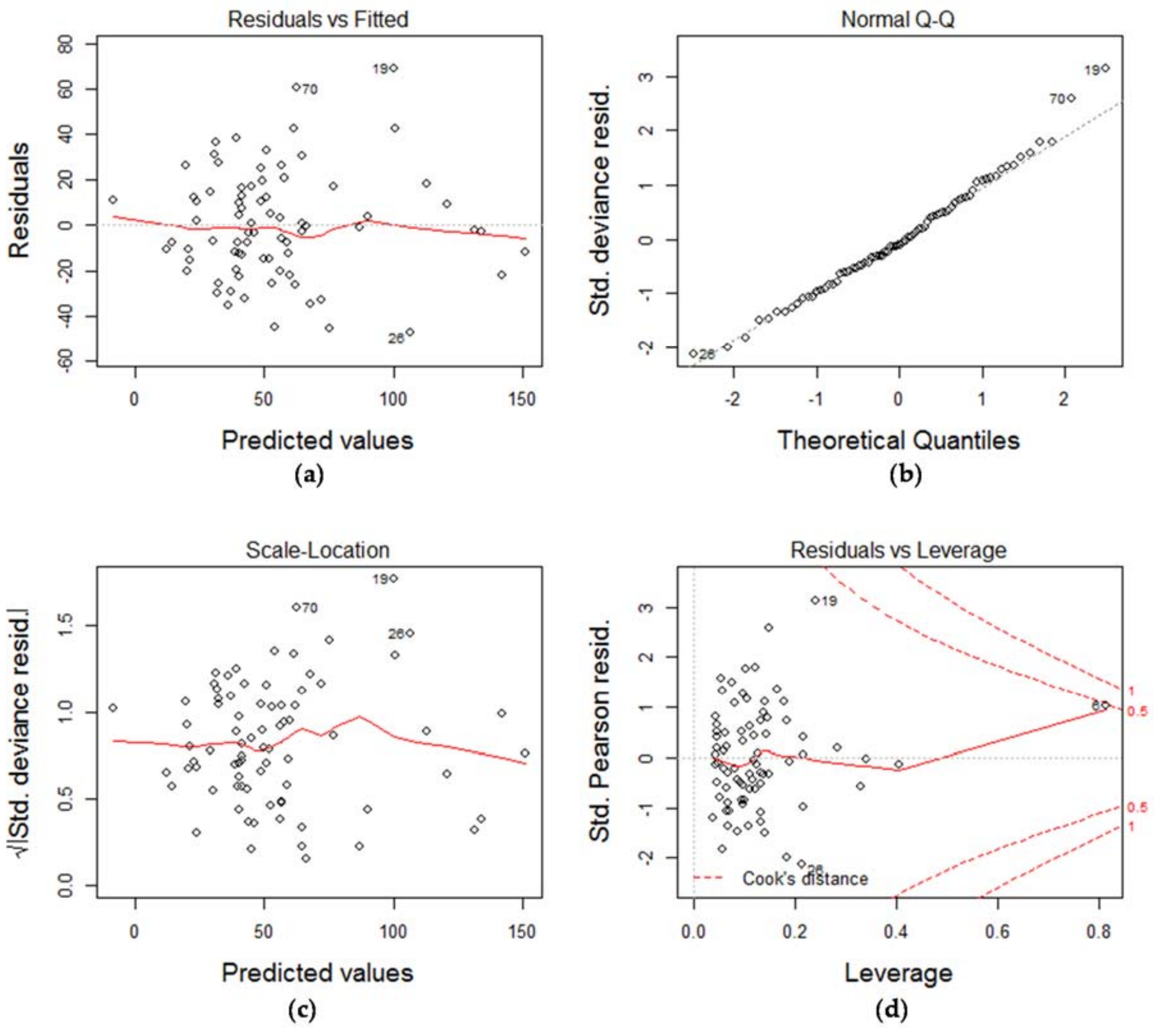


Figure S14. Residuals plots of model based on folds training set 1 + 3 with lowest AIC, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

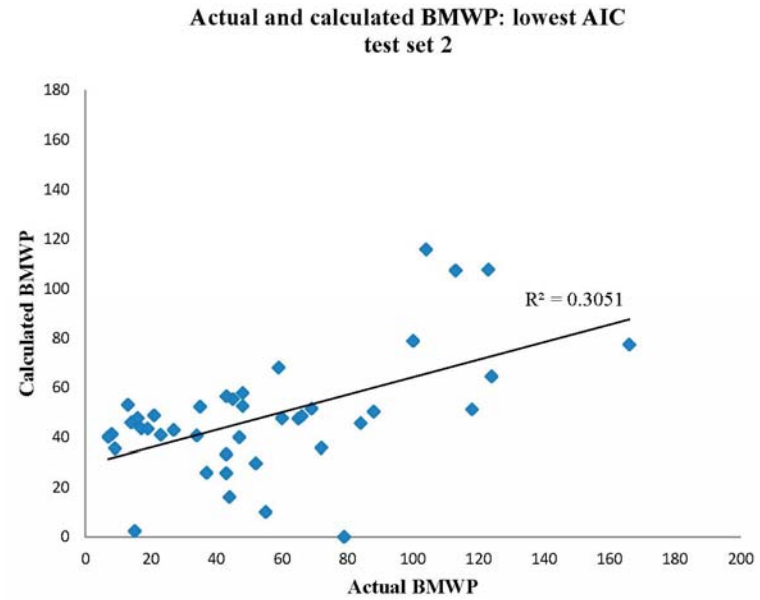


Figure S15. Validation of model based on folds test set 2 with lowest AIC.

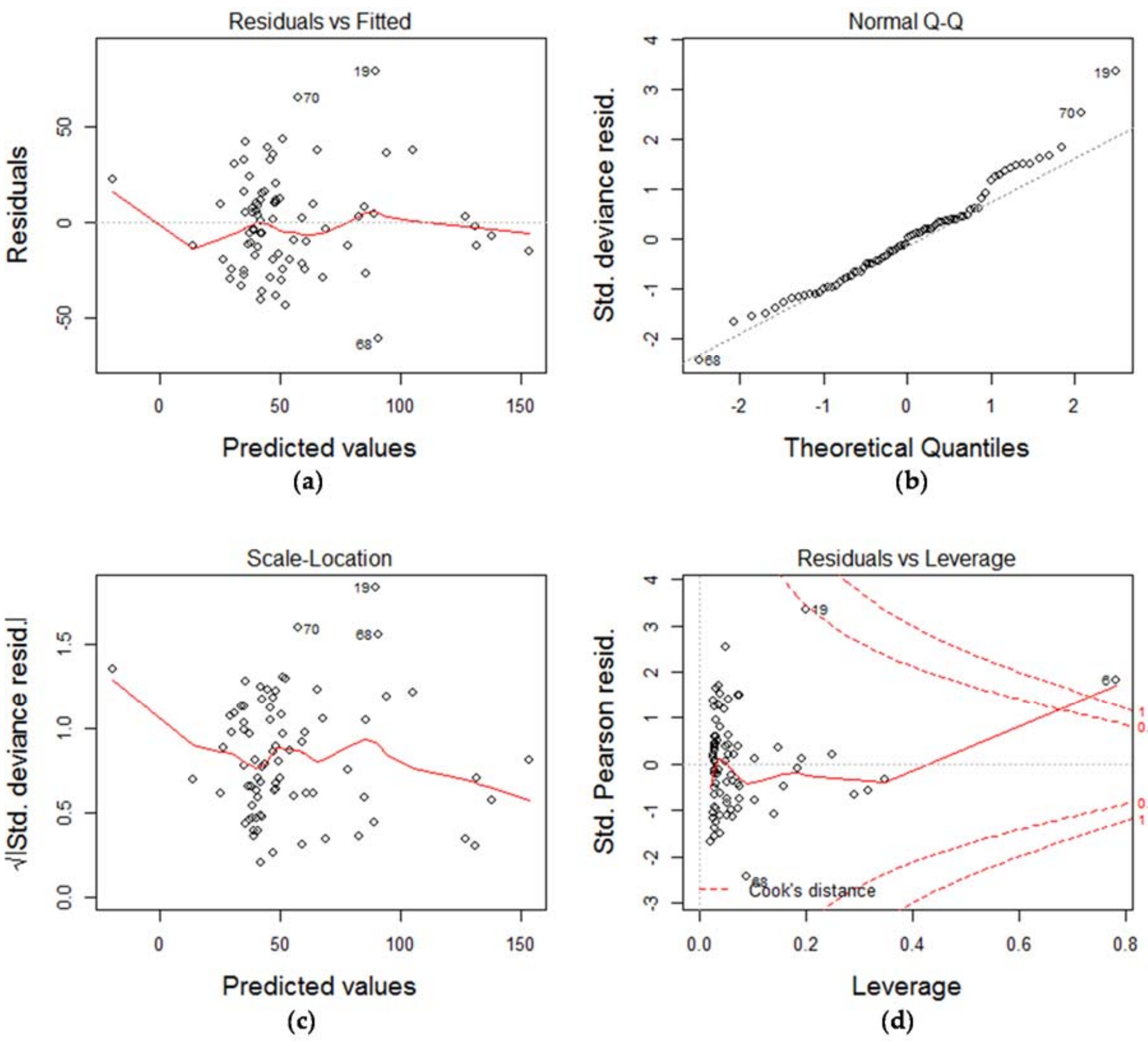


Figure S16. Residuals plots of model based on folds training set 1 + 3 with input variables significant at $p < 0.1$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

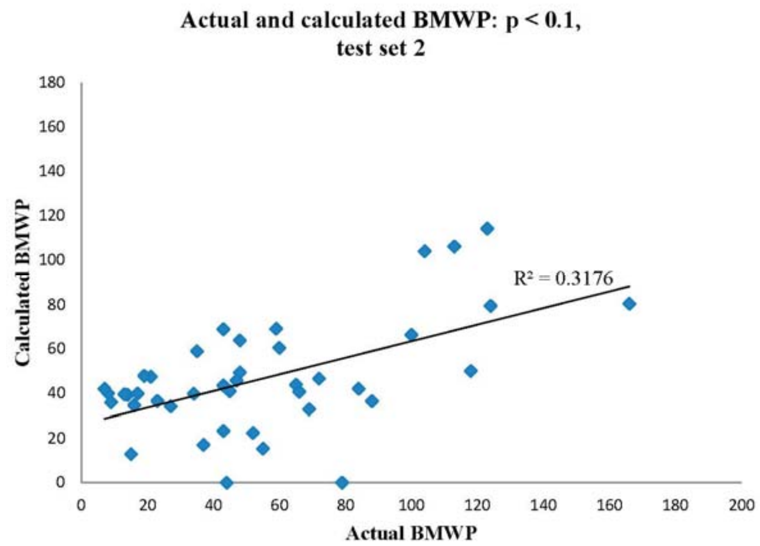


Figure S17. Validation of model based on folds test set 2 with input variables significant at $p < 0.1$.

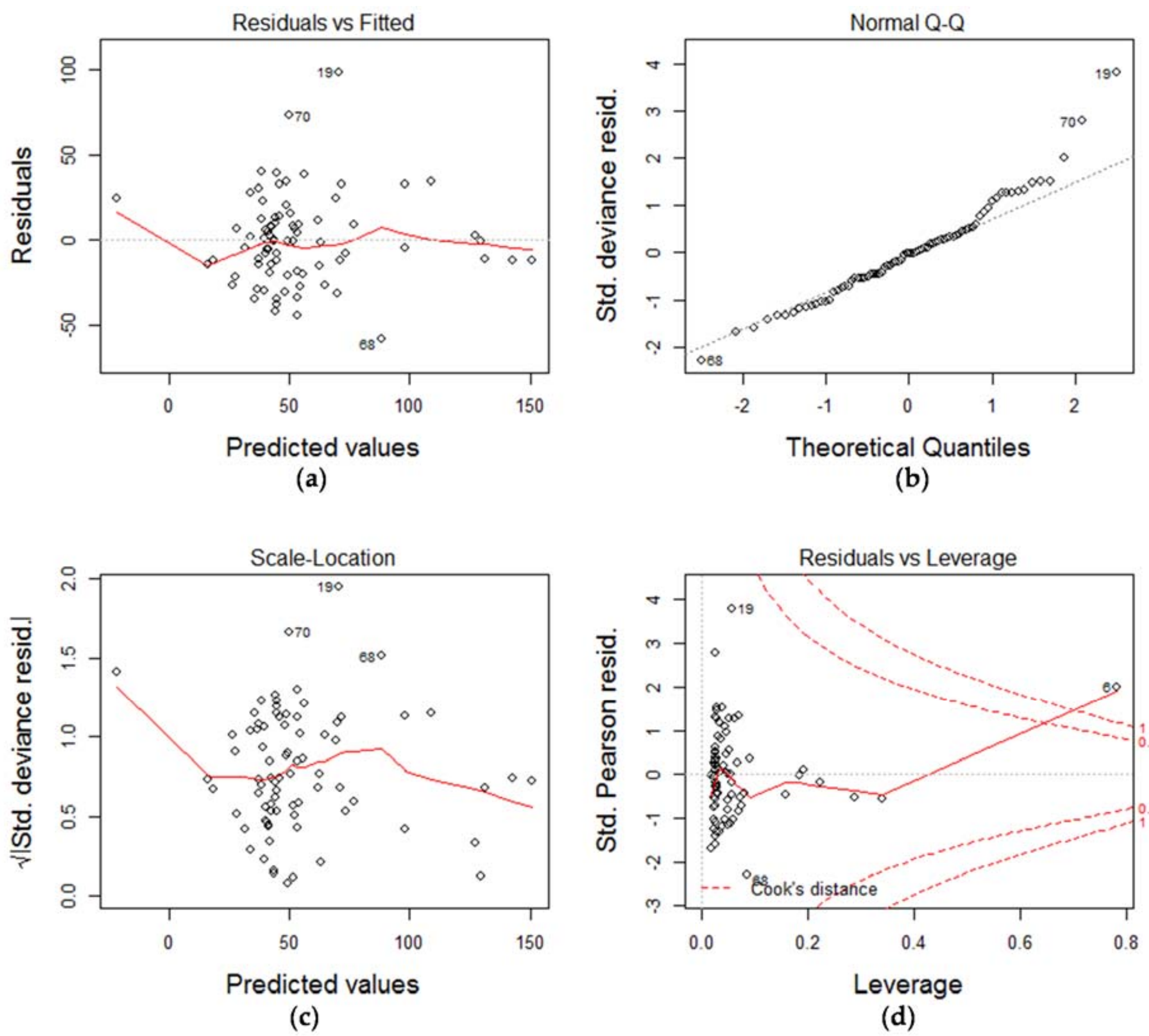


Figure S18. Residuals plots of model based on folds training set 1 + 3 with input variables significant at $p < 0.05$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

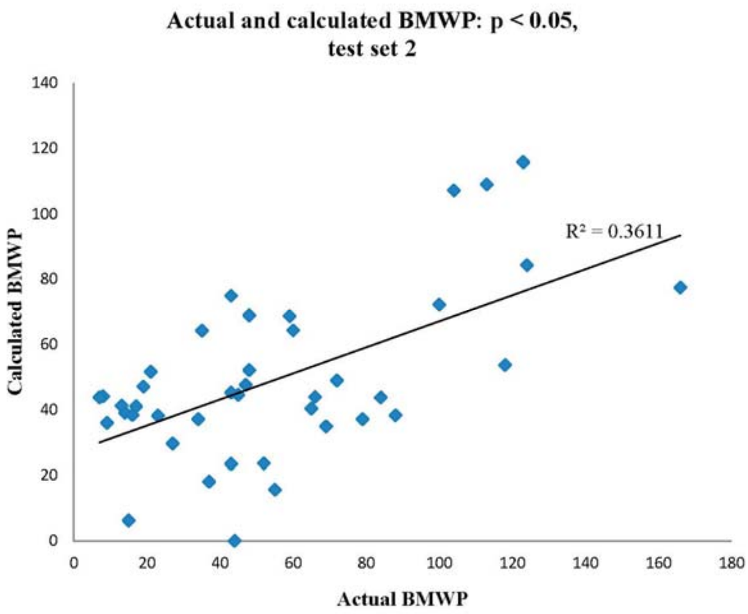


Figure S19. Validation of model based on folds test set 2 with input variables significant at $p < 0.05$.

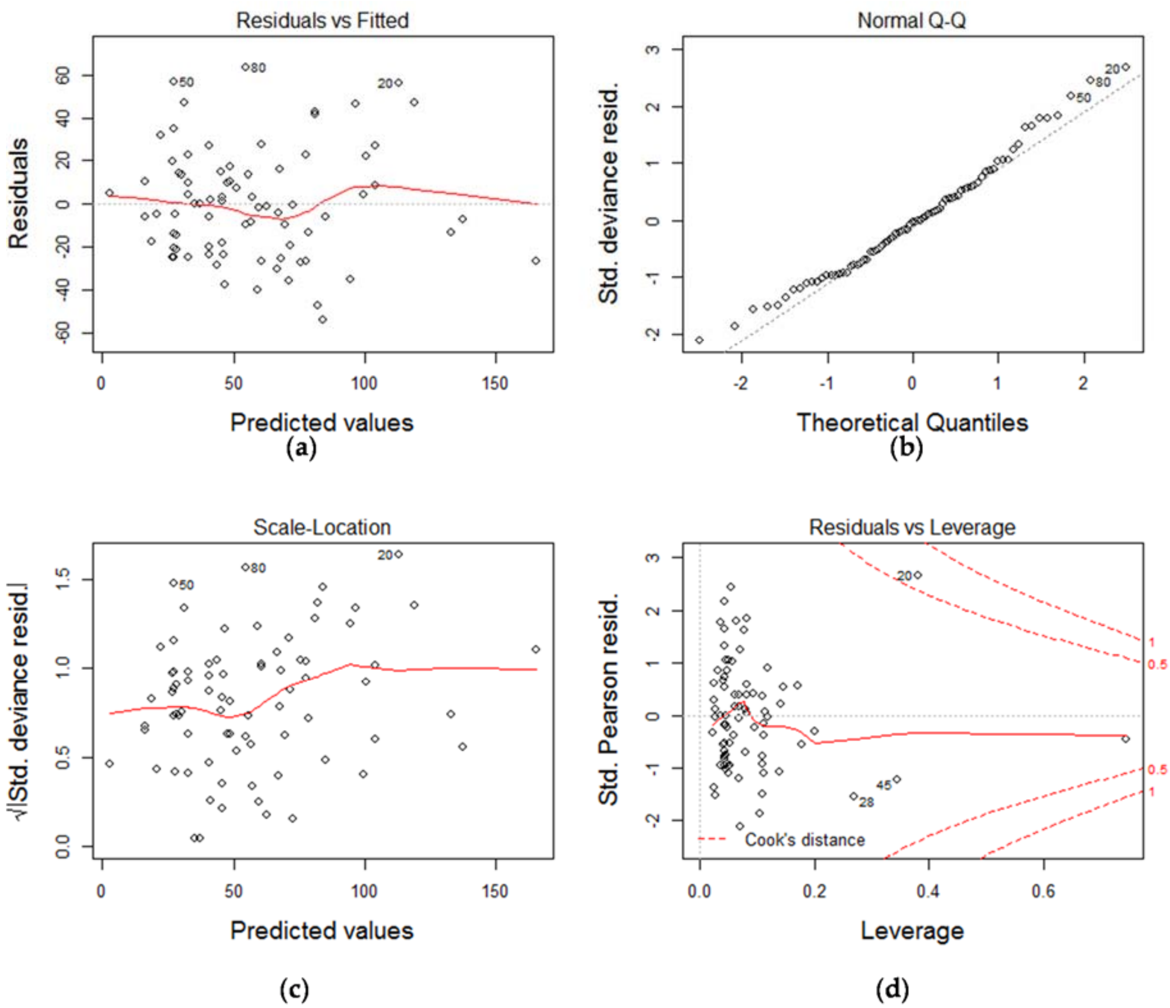


Figure S20. Residuals plots of model based on folds training set 2 + 3 with lowest AIC and input variables significant at $p < 0.1$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

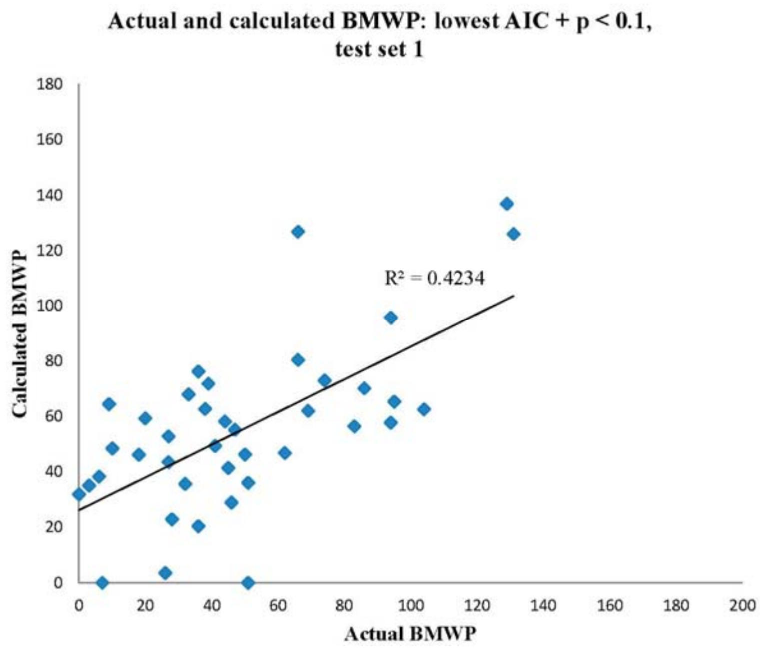


Figure S21. Validation of model based on folds test set 1 with lowest AIC and input variables significant at $p < 0.1$.

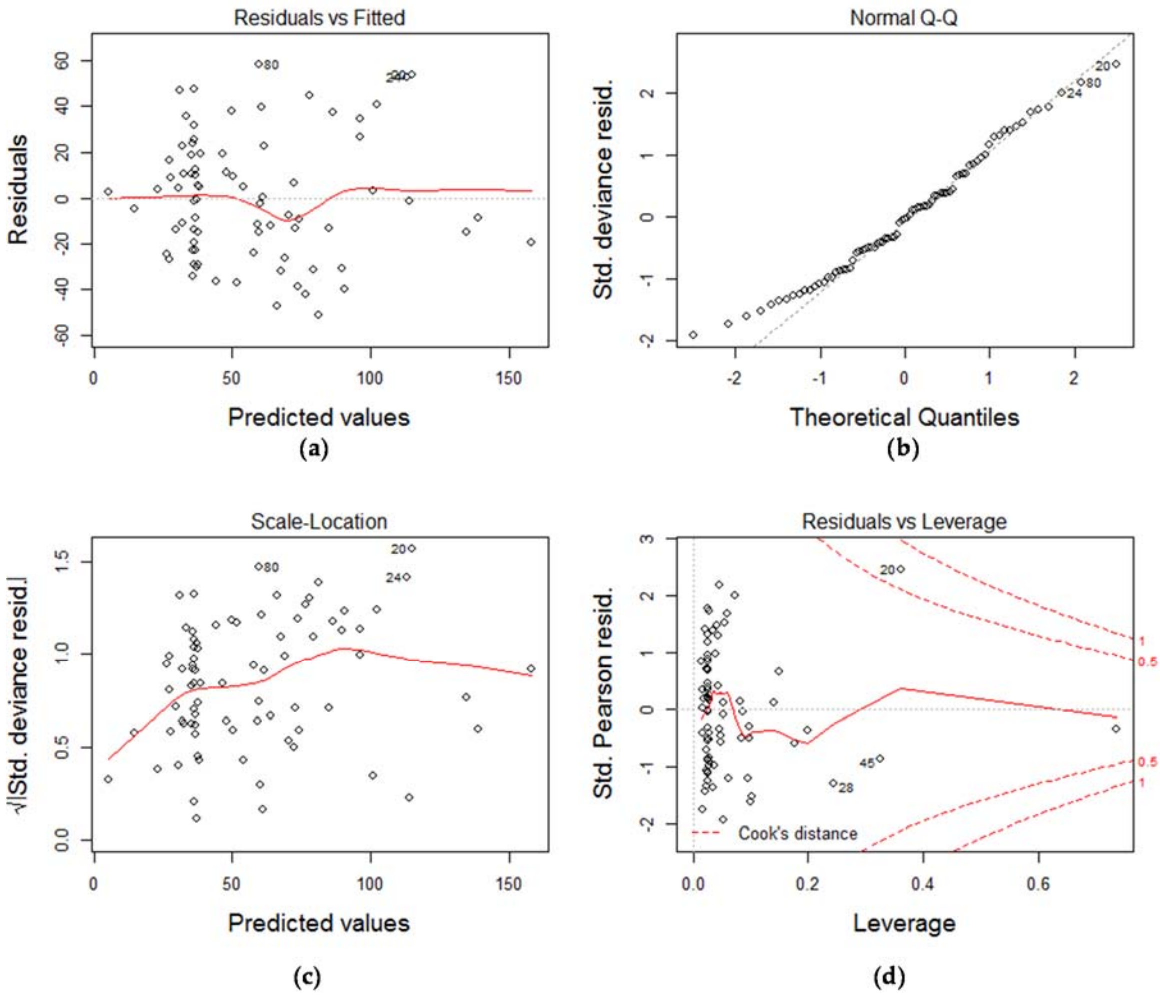


Figure S22. Residuals plots of model based on folds training set 2 + 3 with input variables significant at $p < 0.05$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

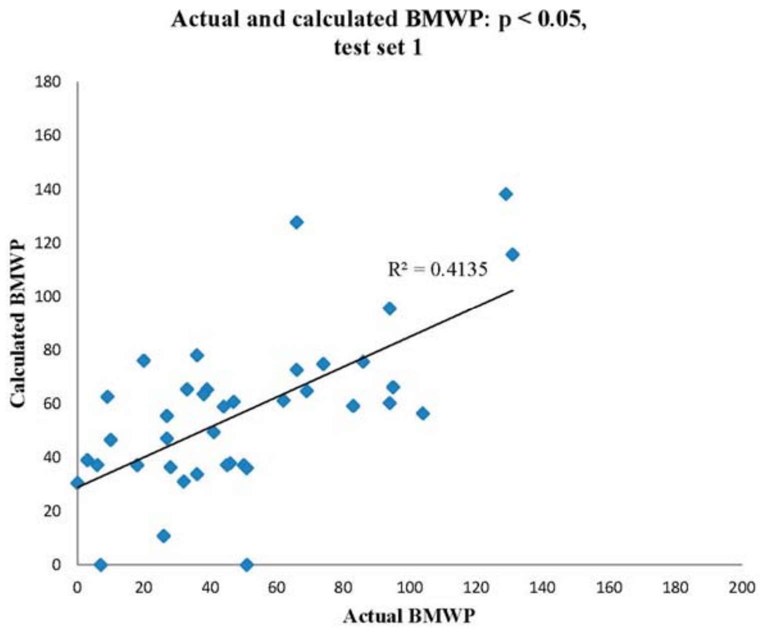


Figure S23. Validation of model based on folds test set 1 with input variables significant at $p < 0.05$.

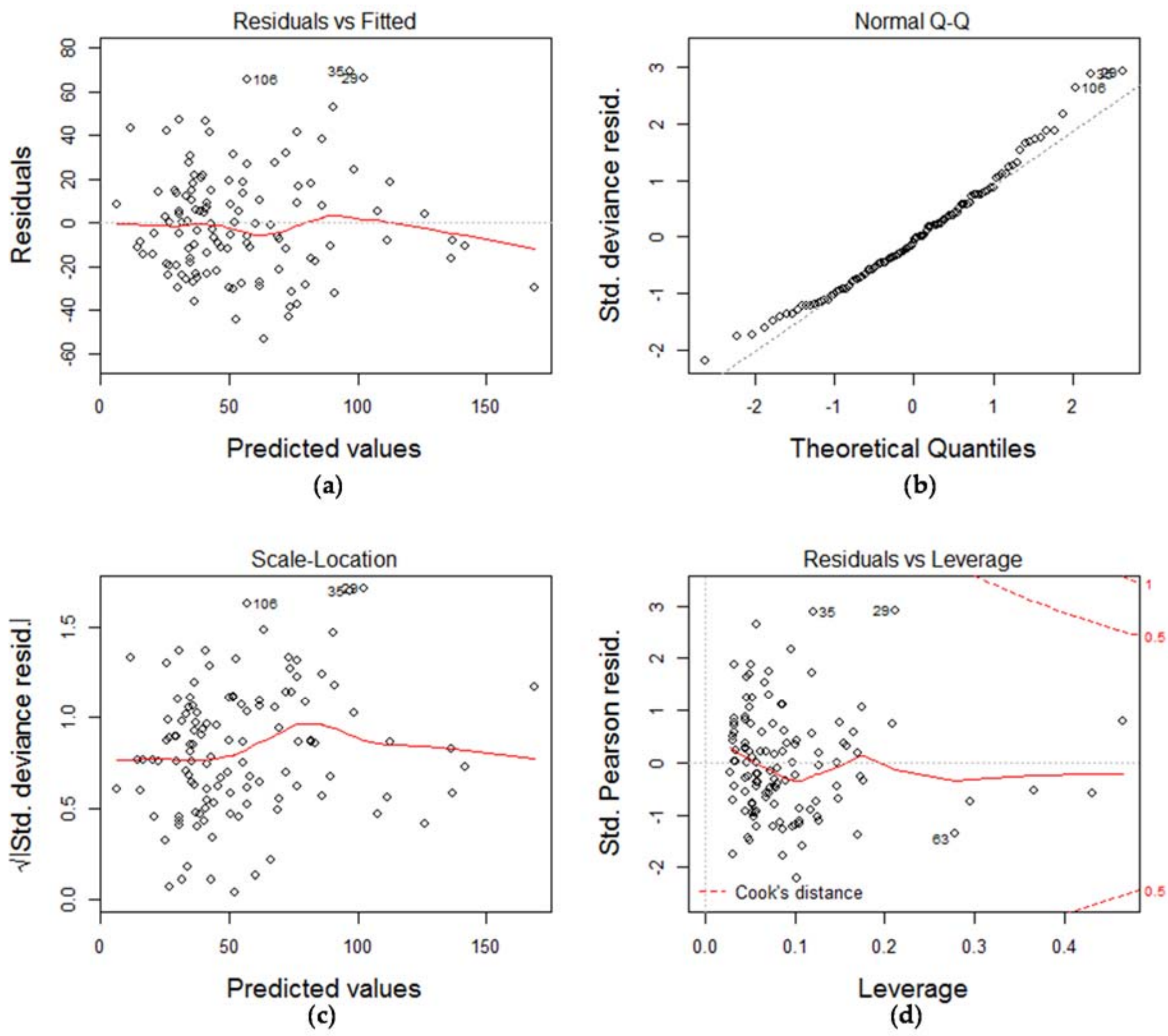


Figure S24. Residuals plots of model with complete data set and lowest AIC, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

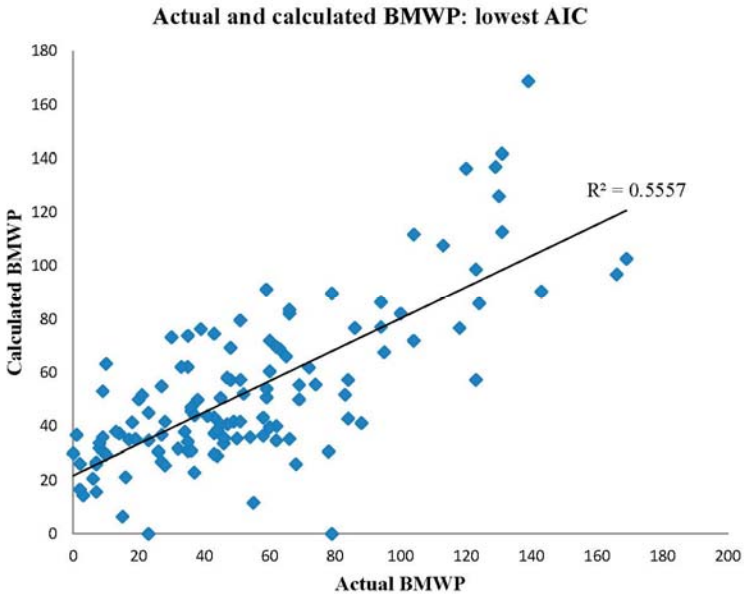


Figure S25. Validation of model with complete data set and lowest AIC.

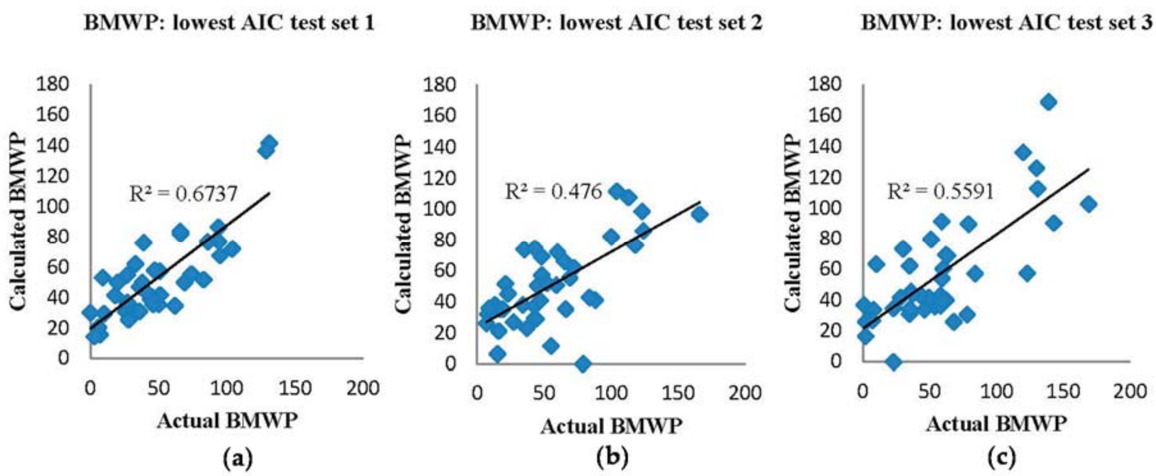


Figure S26. Validation of model with complete data set and lowest AIC on three folds, (**a**) for test set 1; (**b**) for test set 2; (**c**) for test set 3.

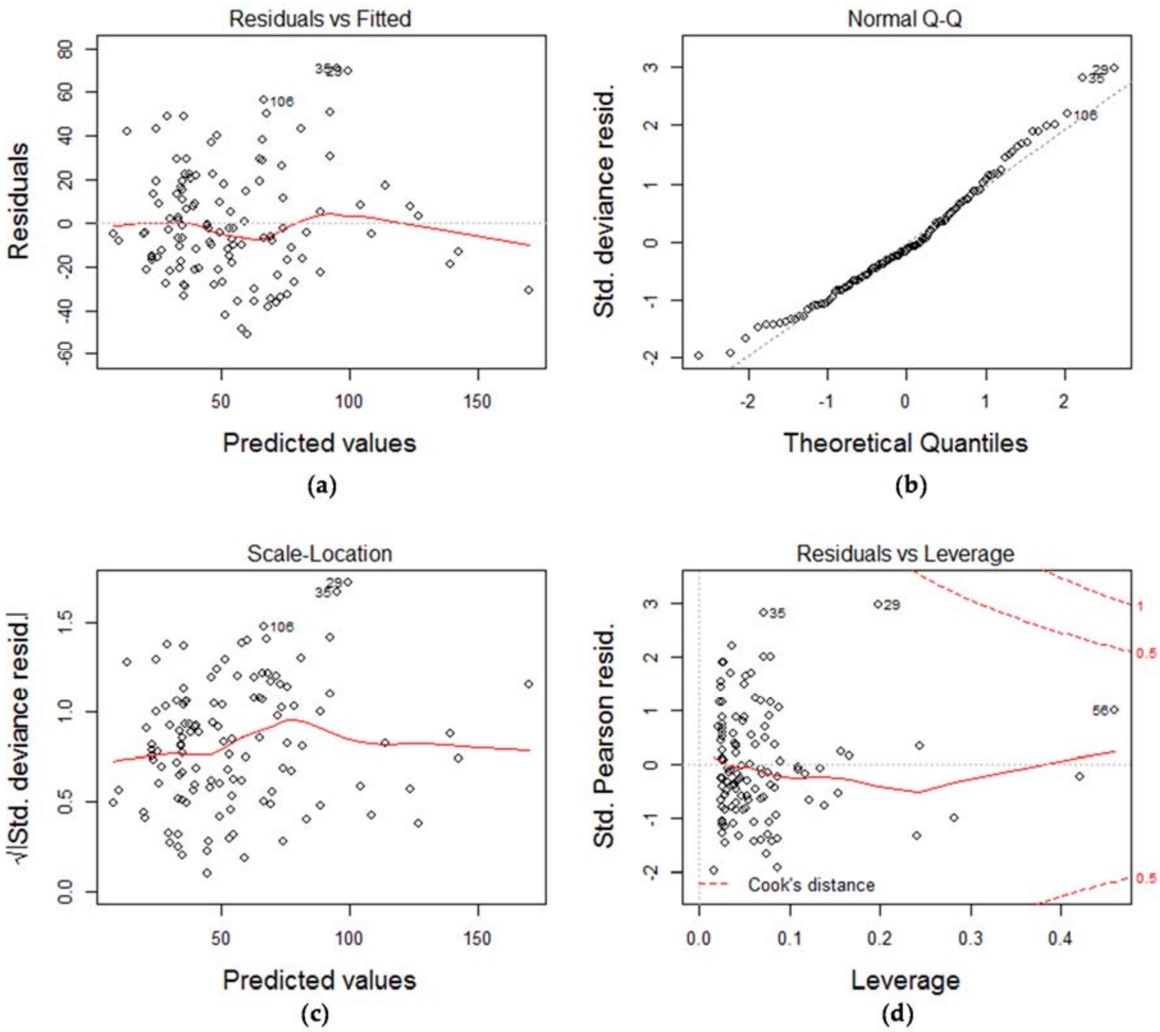


Figure S27. Residuals plots of model with complete data set and input variables significant at $p < 0.1$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

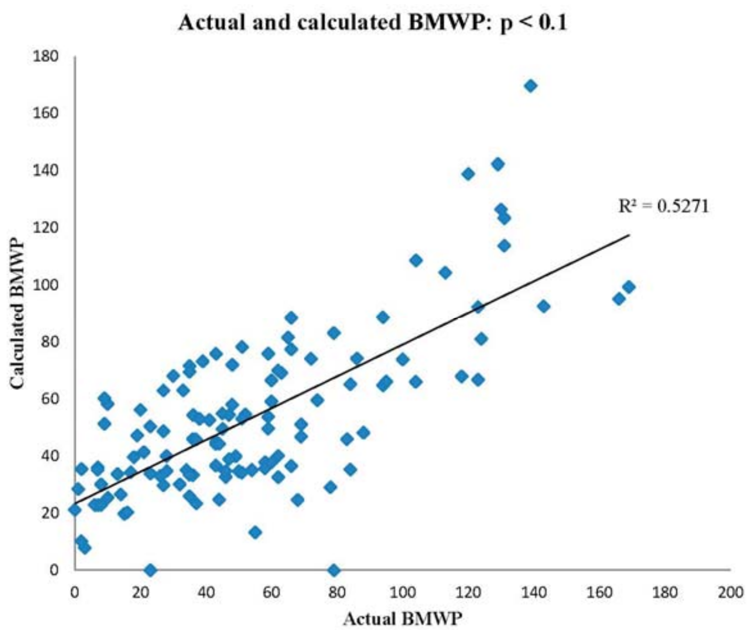


Figure S28. Validation of model with complete data set and input variables significant at $p < 0.1$.

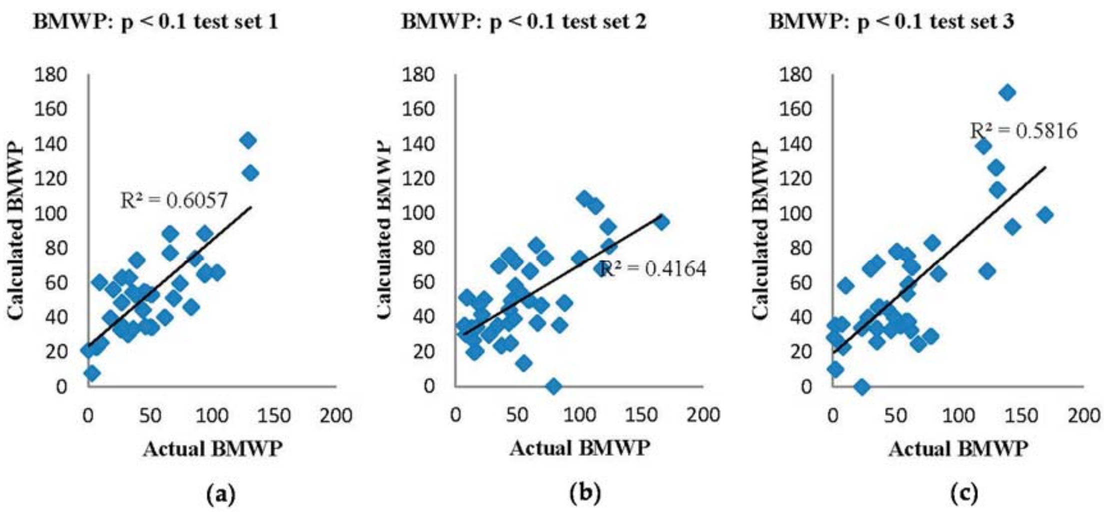


Figure S29. Validation of model with complete data set and input variables significant at $p < 0.1$ on three folds, (**a**) for test set 1; (**b**) for test set 2; (**c**) for test set 3.

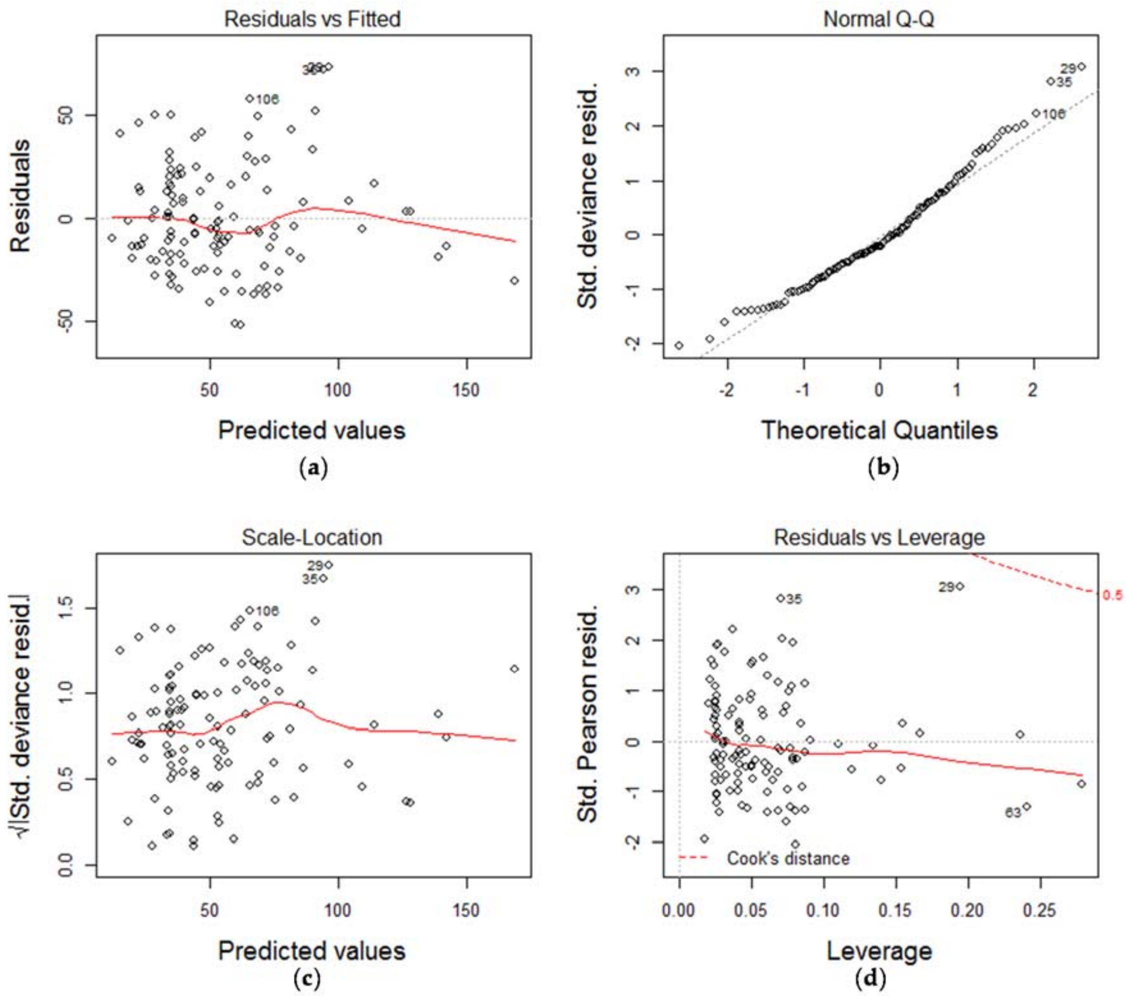


Figure S30. Residuals plots of model with complete data set and input variables significant at $p < 0.05$, (**a**) residuals versus fitted values; (**b**) QQ-plot for normality; (**c**) scaled residuals versus fitted values; (**d**) standardized residuals versus leverage.

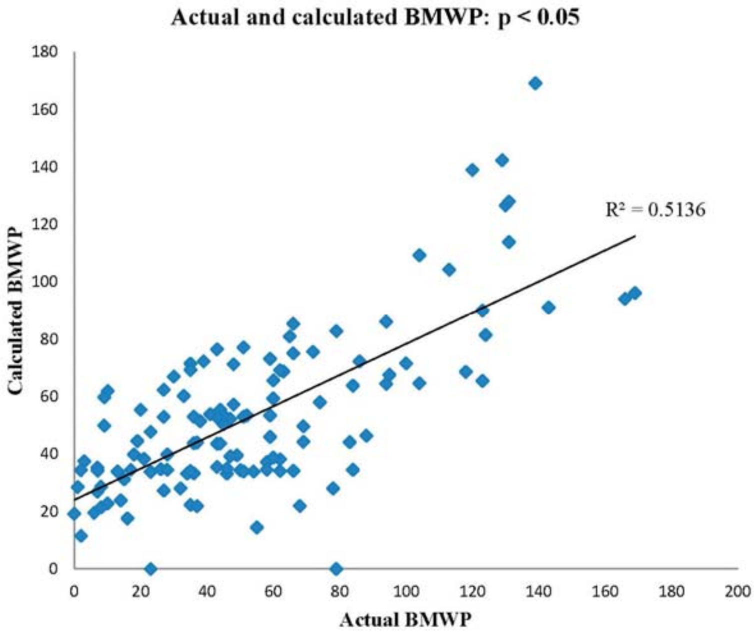


Figure S31. Validation of model with complete data set and input variables significant at $p < 0.05$.

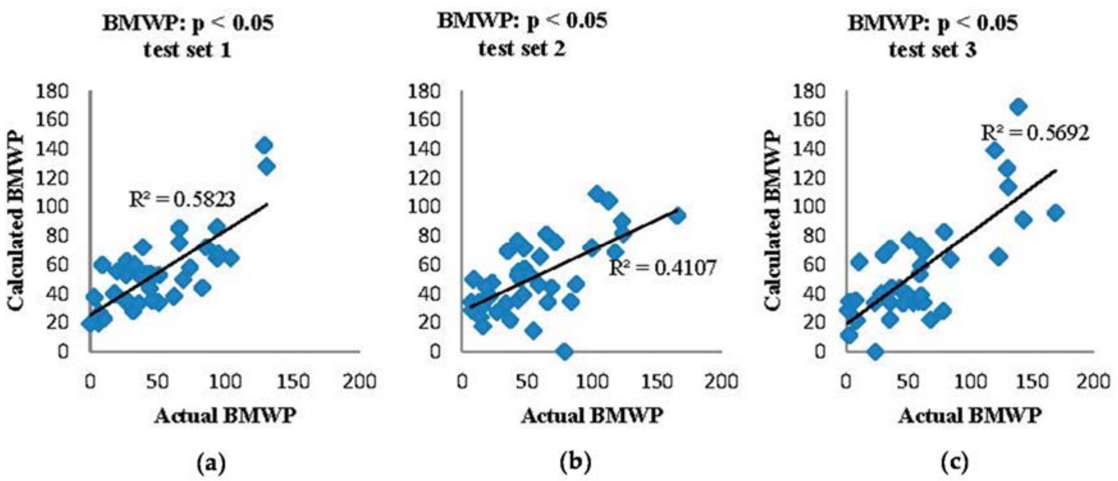


Figure S32. Validation of model with complete data set and input variables significant at $p < 0.05$ on three folds, (**a**) for test set 1; (**b**) for test set 2; (**c**) for test set 3.

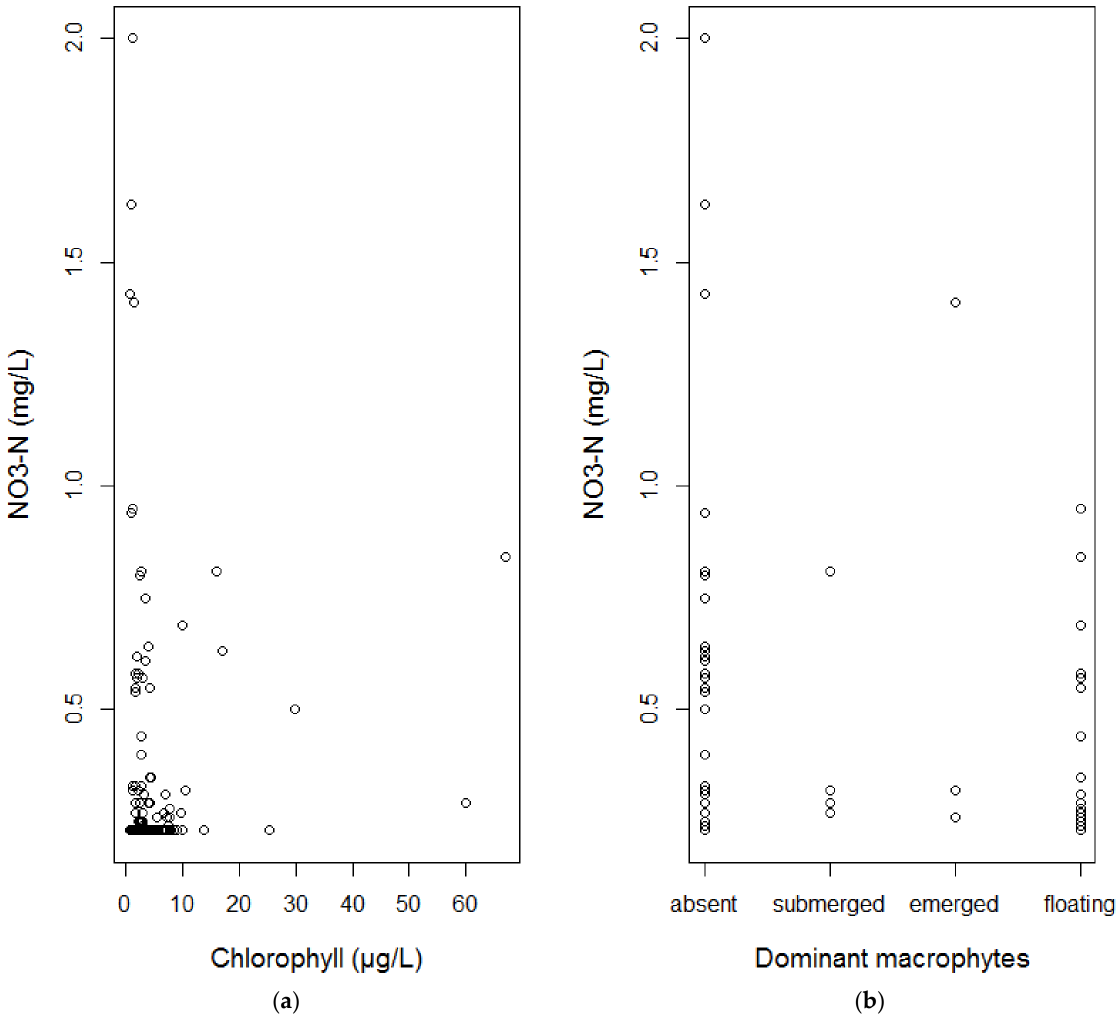


Figure S33. Relationship between nitrate-N and chlorophyll (**a**), and between nitrate-N and dominant macrophytes (**b**).

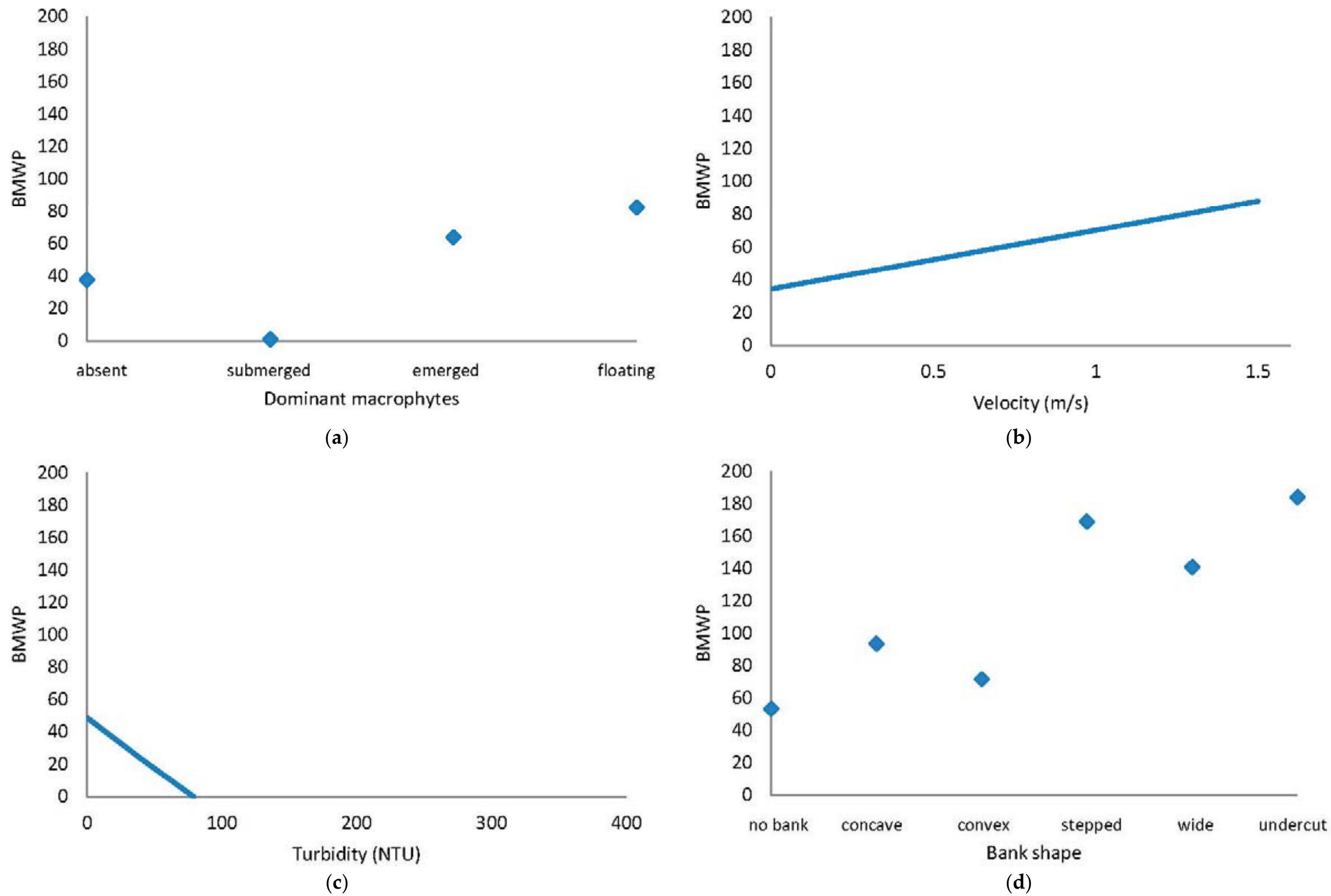
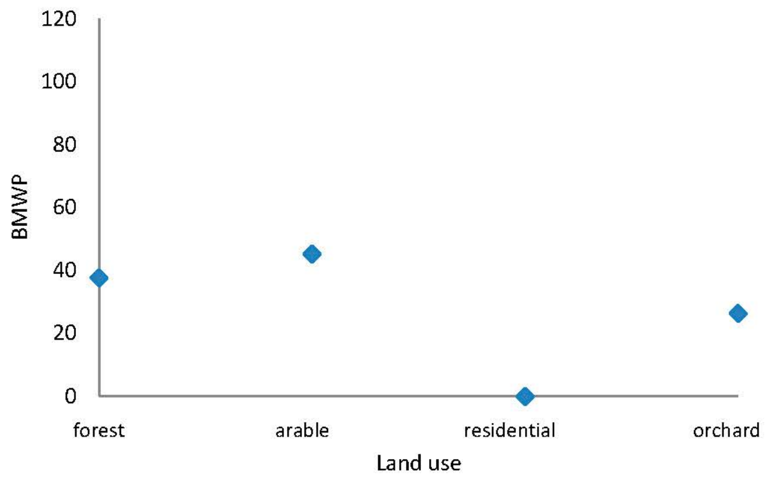
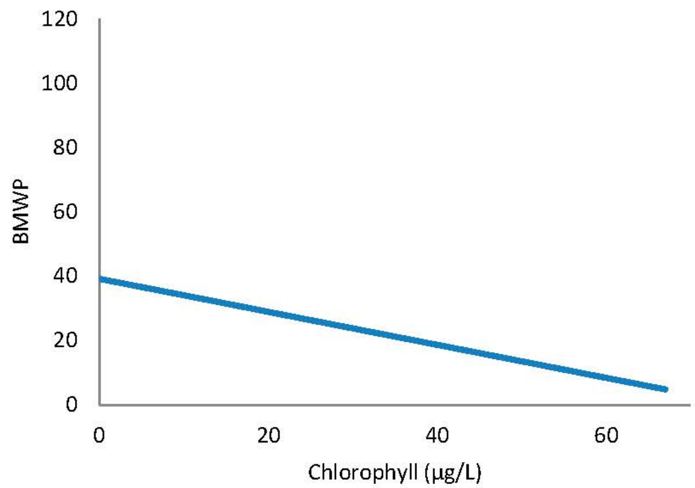


Figure S34. The impact of changing the presence of main macrophytes (**a**); velocity (**b**), turbidity (**c**) and bank shape (**d**) on the biotic integrity expressed as BMWP-Col, *p*-value 0.013, 0.015, 0.05 and 0.036, respectively. The values used in the analysis were based on Table S8, showing the median, minimum and maximum values for sensitivity analysis.



(**a**)



(**b**)

Figure S35. The impact of changing the land use (**a**) and chlorophyll concentration (**b**) on the biotic integrity expressed as BMWP-Col, *p*-value 0.048 and 0.064, respectively. The values used in the analysis were based on Table S8, showing the median, minimum and maximum values for sensitivity analysis.

Table S1. List of families encountered in the Guayas river basin with tolerance scores based on Alvarez, 2005; number of presences in the samples and functional feeding group (FFG).

Family	BMWP-Col (Alvarez, 2005)	# Presence	FFG
Acari	0	56	predator
Aeshnidae	6	11	predator
Ampullariidae	9	6	scraper
Ancylidae	6	13	scraper
Baetidae	7	64	scraper
Belostomatidae	5	6	predator
Blepharoceridae	10	2	scraper
Caenidae	7	12	collector-gatherer
Calamoceratidae	10	2	shredder
Calopterygidae	7	14	predator
Cambaridae	0	7	collector-gatherer
Ceratopogonidae	3	22	collector-gatherer
Chaoboridae	0	2	predator
Chironomidae	2	100	collector-gatherer
Chordodidae	10	1	parasite
Coenagrionidae	7	50	predator
Corbiculidae	0	22	collector-filterer
Corixidae	7	28	predator
Corydalidae	6	11	predator
Coryphoridae	0	7	scraper
Crambidae	5	13	shredder
Culicidae	2	14	collector-gatherer
Dixidae	7	2	collector-gatherer
Dryopidae	7	2	scraper
Dugesiidae	0	33	parasite
Dytiscidae	9	13	predator
Elmidae	6	20	scraper
Empididae	4	4	collector-gatherer
Gerridae	8	26	predator
Glossiphoniidae	3	28	predator
Glossosomatidae	7	2	scraper
Gomphidae	10	17	predator
Gyrinidae	9	1	predator
Hebridae	8	1	predator
Helicopsychidae	8	14	scraper
Heteroceridae	0	1	scraper
Hyallelidae	7	26	shredder
Hydrobiidae	8	7	scraper
Hydrobioscidae	9	3	predator
Hydrometridae	3	1	predator
Hydrophilidae	3	13	predator
Hydropsychidae	7	31	collector-filterer
Hydroptilidae	7	12	collector-gatherer
Lampyridae	10	1	predator
Leptoceridae	8	27	shredder
Leptohyphidae	7	52	collector-gatherer
Leptophlebiidae	9	30	collector-gatherer
Libellulidae	6	55	predator

Limoniidae	3	14	collector-gatherer
Lumbriculidae	0	2	collector-gatherer
Lymnaeidae	4	10	scraper
Macroveliidae	0	4	predator
Megapodagrionidae	6	6	predator
Mesoveliidae	5	15	predator
Mysidae	0	1	scraper
Naucoridae	7	33	predator
Nereidae	0	1	-
Noteridae	4	3	predator
Notonectidae	7	17	predator
Ocypodidae	0	2	shredder
Odontoceridae	10	2	collector-gatherer
Oligoneuriidae	10	1	collector-filterer
Palaemonidae	8	3	scraper
Perlidae	10	12	predator
Philopotamidae	9	17	collector-filterer
Physidae	3	9	scraper
Planorbidae	5	8	scraper
Platystictidae	0	11	predator
Pleidae	8	7	predator
Polycentropodidae	9	6	collector-gatherer
Polymitarcidae	9	1	collector-gatherer
Psephenidae	10	17	scraper
Ptilodactylidae	10	4	shredder
Scirtidae	7	3	collector-gatherer
Simuliidae	8	8	collector-filterer
Sphaeriidae	4	1	collector-filterer
Staphylinidae	6	3	predator
Stratiomyidae	4	7	collector-gatherer
Tabanidae	5	6	predator
Thiaridae	5	36	scraper
Trichodactylidae	0	1	shredder
Tubificidae	1	29	collector-gatherer
Veliidae	8	40	predator

Table S2. Adapted habitat disturbance criteria and scoring list.

Disturbance		Score = 1	Score = 2	Score = 3	Score = 4
Habitat alteration	Grazing	Minimal grazing	Moderate grazing	Intensive grazing	
	Vegetation removal	<10% vegetation removal	10%–50% of vegetation removal	>50% vegetation removal	
	Tree plantation	No tree plantation or plantation at >50 m	Tree plantation at <50 m but not in the wetland	Tree plantation in the wetland	
	Grading	no grading	grading near the wetland	grading within the wetland	
	Filling	No filling	Filling near the wetland	Filling in the wetland	
Land scape	Land use	forested	pasture	arable	residential
	Farming	No farming	Less intensive farming	Intensive farming	
	Soil mining	No soil mining	Soil mining >50 m	Soil mining in the wetland or <50 m	
	Wastewater discharge	No wastewater discharge into the river	Treated wastewater discharge into the river	Untreated wastewater discharge into the river	
Hydrological modification	Water inlet	natural source	berm	dam	
	Draining and water abstraction	no draining nor abstraction	Draining nearby <50 m	Draining in the wetland	

Table S3. Variables' selection for three folds: showing variable with the highest *p*-value in the model together with the AIC of each model.

Model with Training Set 1 + 2	Model with Training Set 1 + 3	Model with Training Set 2 + 3
Bed compaction, $p = 0.97954$, AIC = 754.92	Valley form, $p = 0.90422$, AIC = 757.05	Bed compaction, $p = 0.95311$, AIC = 773.14
Bank profile, $p = 0.93177$, AIC = 752.92	Shading, $p = 0.8122$, AIC = 755.07	Bank slope, $p = 0.93633$, AIC = 771.15
Sludge layer, $p = 0.83762$, AIC = 750.93	Turbidity, $p = 0.753$, AIC = 753.16	NO2, $p = 0.92297$, AIC = 769.16
Valley form, $p = 0.8139$, AIC = 748.99	Bank profile, $p = 0.73639$, AIC = 751.3	Sludge layer, $p = 0.86893$, AIC = 767.17
Branch, $p = 0.68334$, AIC = 747.07	Sediment angularity, $p = 0.55891$, AIC = 749.46	Velocity, $p = 0.75993$, AIC = 765.21
Bank shape, $p = 0.63615$, AIC = 745.3	Variation in flow, $p = 0.50215$, AIC = 747.93	Chlorophyll, $p = 0.782108$, AIC = 763.34
DO, $p = 0.57422$, AIC = 743.6	Main macrophytes, $p = 0.4821$, AIC = 746.54	Branch, $p = 0.672055$, AIC = 761.44
Turbidity, $p = 0.57365$, AIC = 742.02	Sludge layer, $p = 0.531783$, AIC = 745.2	Erosion, $p = 0.528448$, AIC = 759.68
NO2, $p = 0.50653$, AIC = 740.43	Width variation, $p = 0.521583$, AIC = 743.71	Channel form, $p = 0.485696$, AIC = 758.2
Channel form, $p = 0.43673$, AIC = 739	Erosion, $p = 0.3957$, AIC = 742.24	Bank profile, $p = 0.553846$, AIC = 756.82
Width variation, $p = 0.31746$, AIC = 737.77	DO, $p = 0.294835$, AIC = 741.15	Land use, $p = 0.359555$, AIC = 755.26

Twigs, p = 0.231301, AIC = 737.01	Bed compaction, p = 0.3629, AIC = 740.52	Main macrophytes, p = 0.336407, AIC = 754.31
Shading, p = 0.204631, AIC = 736.76	Channel form, 0.4092, AIC = 739.54	Width variation, p = 0.336441, AIC = 753.44
Bank slope, p = 0.18118, AIC = 736.7	NO_2, p = 0.383788, AIC = 738.36	Variation in flow, p = 0.184036, AIC = 752.55
Velocity, p = 0.099362, AIC = 736.83	Twigs, p = 0.31435, AIC = 737.27	DO, p = 0.233515, AIC = 752.65
Main land use, p = 0.160108, AIC = 738.01	Branch, p = 0.141006, AIC = 736.45	Bank shape, p = 0.224147, AIC = 752.31
Chlorophyll, p = 0.043185, AIC = 738.28	Logs, p = 0.112273, AIC = 736.96	Logs, p = 0.147498, AIC = 752.01
Main macrophytes, p = 0.040888, AIC = 740.93	Main land use, p = 0.12452, AIC = 737.83	Shading, p = 0.25308, AIC = 752.4
NO3, p = 0.109585, AIC = 743.61	Bank slope, p = 0.10978, AIC = 738.48	Twigs, p = 0.06822, AIC = 751.86
Logs, p = 0.131943, AIC = 755.67	NO_3, p = 0.05478, AIC = 739.31	Valley form, p = 0.103873, AIC = 753.54
Variation in flow, 0.046539, AIC = 756.14	Chlorophyll, p = 0.022016, AIC = 749.99	NO_3, p = 0.006893, AIC = 754.42

Table S4. Ranking of importance of input variables in the models with 3-folds cross validation, based on the p-values (the p-values are given between brackets).

Variables	**Variables' Ranking**										
	First	**Second**	**Third**	**Fourth**	**Fifth**	**Sixth**	**Seventh**	**Eighth**	**Ninth**	**Tenth**	**Eleventh**
Training set 1 + 2											
Lowest AIC	Elevation (0.001)	Main macrophytes (0.013)	NO_3^--N (0.024)	Angularity (0.027)	Logs (0.044)	Land use (0.048)	Erosion (0.048)	Chlorophyll (0.064)	Flow variation (0.067)	Velocity (0.151)	Bank slope (0.181)
$p < 0.1$	Elevation (0.003)	Main macrophytes (0.012)	Angularity (0.013)	Logs (0.016)	NO_3^--N (0.019)	Erosion (0.048)	Chlorophyll (0.050)	Flow variation (0.067)	Land use (0.085)	Velocity (0.099)	
$p < 0.05$	Elevation (<0.001)	Angularity (0.005)	Flow variation (0.009)	Erosion (0.014)	Logs (0.016)	Main macrophytes (0.017)	NO_3^--N (0.020)	Chlorophyll (0.043)			
Training set 1 + 3											
Lowest AIC	Elevation (<0.001)	Velocity (0.015)	Bank shape (0.036)	Logs (0.042)	Land use (0.063)	NO_3^--N (0.066)	Chlorophyll (0.072)	Bank slope (0.081)	Branch (0.141)		
$p < 0.1$	Elevation (<0.001)	Velocity (0.004)	Bank shape (0.024)	Chlorophyll (0.033)	NO_3^--N (0.055)						
$p < 0.05$	Elevation (<0.001)	Velocity (<0.001)	Bank shape (0.020)	Chlorophyll (0.022)							
Training set 2 + 3											
Lowest AIC	Elevation (<0.001)	Angularity (0.002)	NO_3^--N (0.003)	Turbidity (0.005)	Valley form (0.056)	Twigs (0.068)					
$p < 0.1$	Elevation (<0.001)	Angularity (0.002)	NO_3^--N (0.003)	Turbidity (0.005)	Valley form (0.056)	Twigs (0.068)					
$p < 0.05$	Elevation	Angularity	Turbidity	NO_3^--N (0.007)							

	(<0.001)	(<0.001)	(0.004)								
# occurrence											
Nine times	Elevation										
Three times		Angularity, velocity	Bank shape								
Two times		Main macrophytes	NO_3^--N	Logs, chlorophyll	Logs, NO_3^--N		Chlorophyll	Chlorophyll		Velocity	
One time			Angularity, flow variation, turbidity	Angularity, erosion, turbidity, NO_3^--N	Land use, valley form	Land use, erosion, main macrophytes, NO_3^--N, twigs	Erosion, NO_3^--N	Flow variation, bank slope	Flow variation, land use, branch		Bank slope

Table S5. Predictive performances of models with 3-folds cross validation: showing the number of input variables that construct the models and the correlation coefficient (R^2) values of training and testing sets.

Models	# Input Variables	R^2	
		Training Set	Testing Set
Fold 1 lowest AIC	11	0.57	0.49
Fold 1 $p < 0.1$	10	0.56	0.48
Fold 1 $p < 0.05$	8	0.54	0.45
Fold 2 lowest AIC	9	0.62	0.31
Fold 2 $p < 0.1$	5	0.56	0.32
Fold 2 $p < 0.05$	4	0.54	0.36
Fold 3 lowest AIC	6	0.55	0.42
Fold 3 $p < 0.1$	6	0.55	0.42
Fold 3 $p < 0.05$	4	0.52	0.41
Average lowest AIC	9	0.58	0.41
Average $p < 0.1$	7	0.56	0.41
Average $p < 0.05$	5	0.53	0.41

Table S6. Variables' selection for model with complete data set: showing variable with the highest *p*-value in the model and the AIC of each model.

Variable with Highest *p*-Value	Model's AIC	Remarks
Bank profile, $p = 0.895125$	1131.9	
Sludge layer, $p = 0.89966$	1129.9	
Channel form, $p = 0.882337$	1127.9	
Erosion, $p = 0.781556$	1126	
NO2, $p = 0.753035$	1124	
Branch, $p = 0.730042$	1122.2	
DO, $p = 0.704733$	1120.3	
Variation in width, $p = 0.576093$	1118.5	
Bed compaction, $p = 0.582336$	1116.9	
Bank slope, $p = 0.36968$	1115.2	
Valley form, $p = 0.353982$	1114.1	
Main land use, $p = 0.204995$	1113.1	
Shading, $p = 0.18378$	1113	
Turbidity, $p = 0.184818$	1113	
Logs, $p = 0.112776$	1112.9	Lowest AIC, $R^2 = 0.56$
Main macrophytes, $p = 0.10811$	1113.7	
Velocity, $p = 0.13639$	1114.5	
Chlorophyll, $p = 0.06668$	1115	Significant at $p < 0.1$, $R^2 = 0.53$
Twigs, $p = 0.01639$	1116.6	Significant at $p < 0.05$, $R^2 = 0.51$

Table S7. Variables' ranking of importance based on the *p*-values for models with complete data set, the *p*-values are given between brackets.

	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth
Lowest AIC	Elevation (<0.001)	NO_3^--N (0.004)	Bank shape (0.004)	Angularity (0.013)	Flow variation (0.027)	Chlorophyll (0.056)	Twigs (0.067)	Velocity (0.073)	Main macrophytes (0.087)	Logs (0.113)
$p < 0.1$	Elevation (<0.001)	NO_3^--N (0.003)	Angularity (0.004)	Twigs (0.011)	Flow variation (0.011)	Bank shape (0.013)	Chlorophyll (0.067)			
$p < 0.05$	Elevation (<0.001)	Angularity (0.001)	NO_3^--N (0.005)	Flow variation (0.006)	Bank shape (0.013)	Twigs (0.016)				

Table S8. Median, minimum and maximum values for sensitivity analysis of models based on complete data set.

No	Variables	Unit	Median	Min	Max
1	Chlorophyll	(µg/L)	3.1	0.7	66.8
2	Nitrate-N	(mg/L)	0.2	0.2	2.0
3	Turbidity	(NTU)	3.4	0.0	355.6
4	Velocity	(m/s)	0.2	0.0	1.5
5	Elevation	(m)	82	2	1075
6	Main macrophytes		0	0	3
7	Variation in flow		0	0	4
8	Twigs		0	0	2
9	Logs		0	0	2
10	Bank shape		2	0	5
11	Sediment angularity		5	3	6
12	Valley form		5	2	7
13	Main land use		1	1	4
14	Bank slope		2	0	5
15	Branch		1	0	2
16	Erosion		0	0	2
	BMWP-Col		47	0	169