Supporting Information

Prediction modeling and mapping of groundwater fluoride contamination throughout India

Joel E. Podgorski[1*], Pawan Labhasetwar[2], Dipankar Saha[3] and Michael Berg[1*]

[1] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water,8600 Dübendorf, Switzerland,

[2] Water Technology and Management Division, CSIR-NEERI, Nehru Marg, Nagpur 440020, India

[3] Member (Retired), Central Ground Water Board, Bhujal Bhawan, NH-IV, Faridabad, 121001, India

* corresponding authors, e-mail: joel.podgorski@eawag.ch and michael.berg@eawag.ch
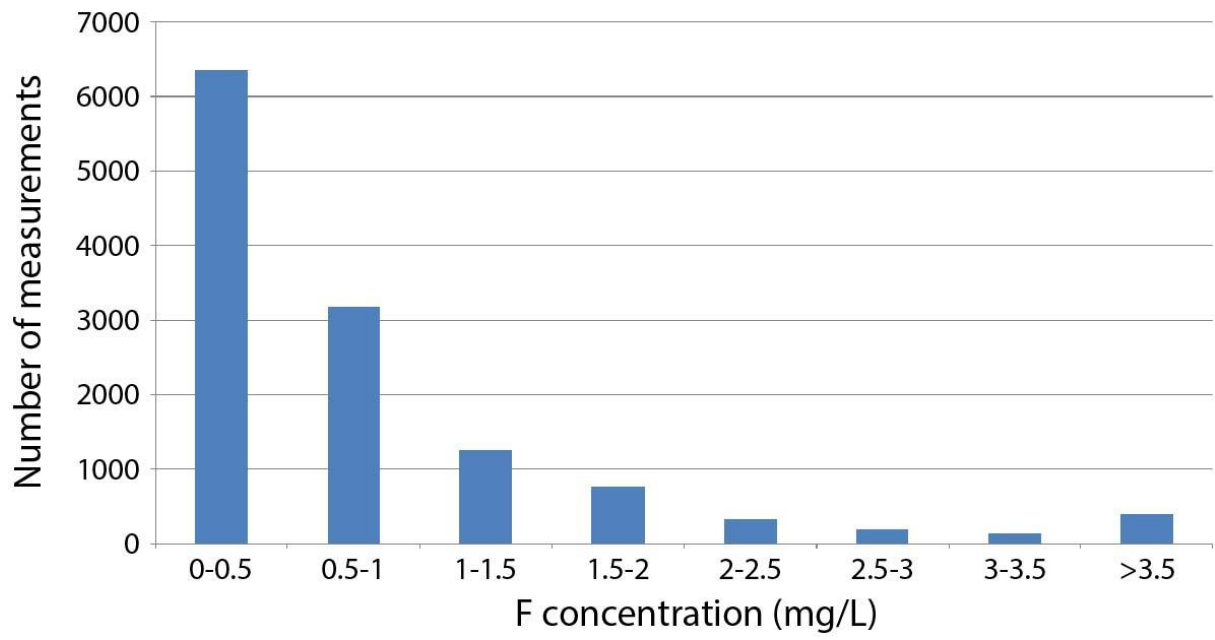
updated 3 August 2018

Contents:

FIGURE S1: Frequency diagram of concentrations of the 12,600 measured concentrations used as input to modeling.
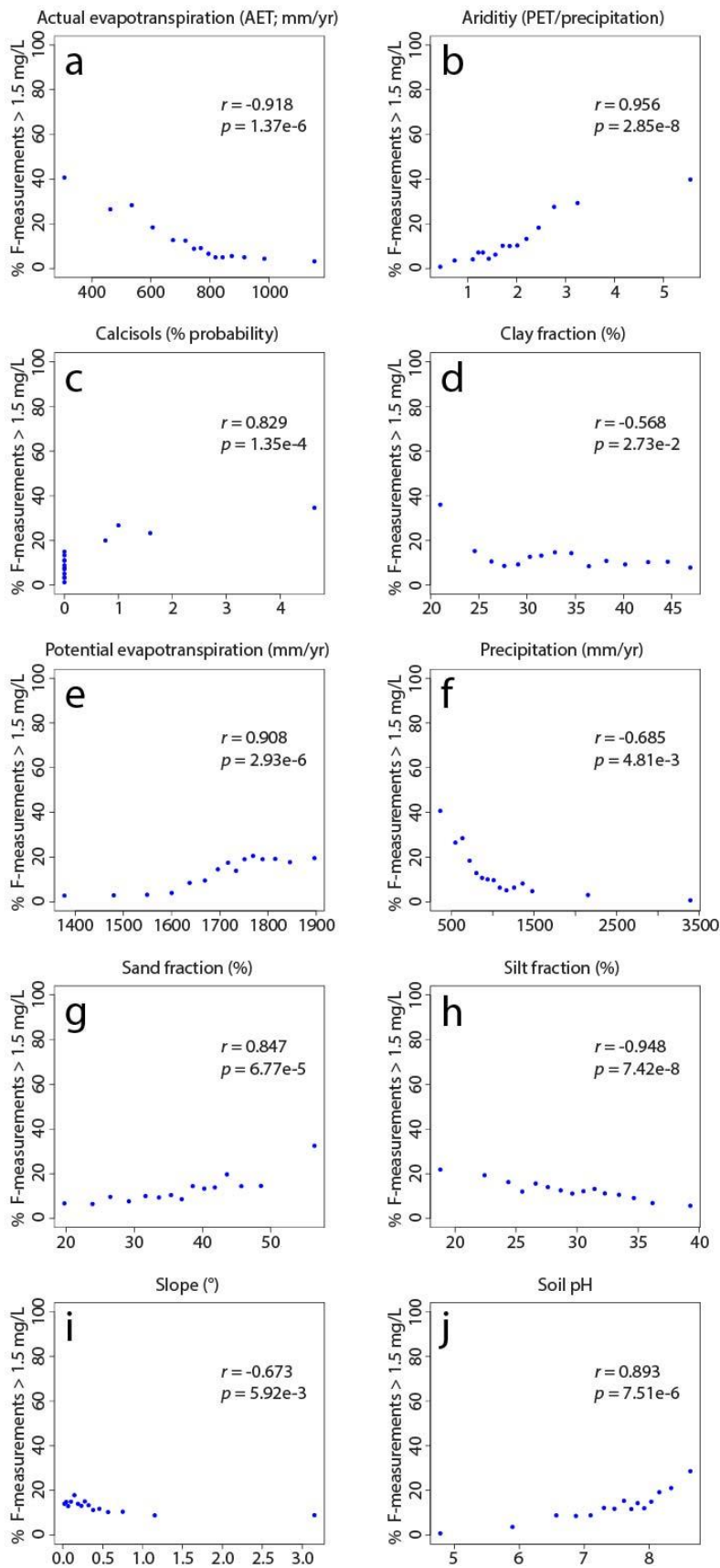
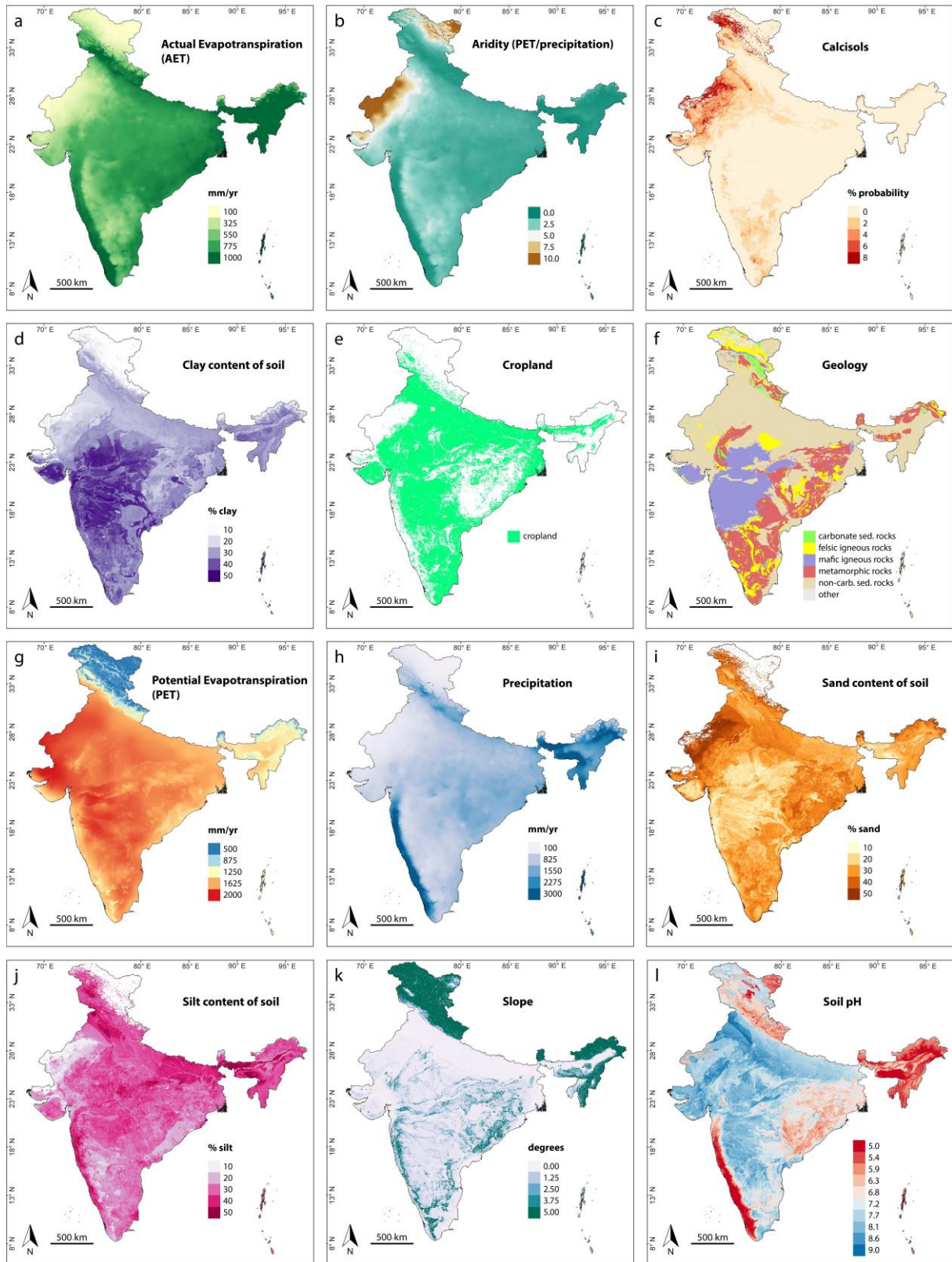**FIGURE S2: Correlation between predictor variables and percentage of measurements greater than 1.5 mg/L.**

FIGURE S3: Predictor variables used in modeling. The data sources are listed in Table 2.

TABLE S1: Correlations among predictor variables. References to data sources are given in Table 2.

	AET	felsic igneous	mafic igneous	aridity	calcisols	clay fract.	slope	cropland	PET	precip.	non-carb. sed.	silt fract.	sand fract.	soil pH
AET	1.0000	0.1339	-0.0800	-0.8607	-0.5701	0.1439	0.1728	-0.1515	-0.6073	0.7697	-0.0708	0.2718	-0.2538	-0.6935
felsic igneous	0.1339	1.0000	-0.1583	-0.0769	-0.0646	-0.0582	0.1167	-0.0709	-0.1008	0.1626	-0.2900	-0.1730	0.1424	-0.2582
mafic igneous	-0.0800	-0.1583	1.0000	-0.0121	-0.1740	0.5980	0.0645	0.0822	0.2580	-0.0534	-0.4635	0.1899	-0.5648	0.1670
aridity	-0.8607	-0.0769	-0.0121	1.0000	0.6456	-0.2564	-0.1622	0.0569	0.5476	-0.6468	0.1456	-0.3581	0.3750	0.5198
calcisols	-0.5701	-0.0646	-0.1740	0.6456	1.0000	-0.4090	-0.0854	-0.0429	0.2005	-0.3404	0.2456	-0.2627	0.4625	0.2726
clay fract.	0.1439	-0.0582	0.5980	-0.2564	-0.4090	1.0000	0.0703	0.0649	0.1324	0.1243	-0.5308	0.1151	-0.8263	0.0219
slope	0.1728	0.1167	0.0645	-0.1622	-0.0854	0.0703	1.0000	-0.1888	-0.2506	0.2382	-0.2118	-0.0324	-0.0345	-0.3411
cropland	-0.1515	-0.0709	0.0822	0.0569	-0.0429	0.0649	-0.1888	1.0000	0.1972	-0.2772	0.0678	0.1730	-0.1412	0.3749
PET	-0.6073	-0.1008	0.2580	0.5476	0.2005	0.1324	-0.2506	0.1972	1.0000	-0.6430	-0.0453	-0.0892	-0.0580	0.6662
precip.	0.7697	0.1626	-0.0534	-0.6468	-0.3404	0.1243	0.2382	-0.2772	-0.6430	1.0000	-0.1238	0.2433	-0.2263	-0.8466
non-carb. sed.	-0.0708	-0.2900	-0.4635	0.1456	0.2456	-0.5308	-0.2118	0.0678	-0.0453	-0.1238	1.0000	0.0927	0.3530	0.2380
silt fract.	0.2718	-0.1730	0.1899	-0.3581	-0.2627	0.1151	-0.0324	0.1730	-0.0892	0.2433	0.0927	1.0000	-0.6340	0.0811
sand fract.	-0.2538	0.1424	-0.5648	0.3750	0.4625	-0.8263	-0.0345	-0.1412	-0.0580	-0.2263	0.3530	-0.6340	1.0000	-0.0609
Soil pH	-0.6935	-0.2582	0.1670	0.5198	0.2726	0.0219	-0.3411	0.3749	0.6662	-0.8466	0.2380	0.0811	-0.0609	1.0000

a

b

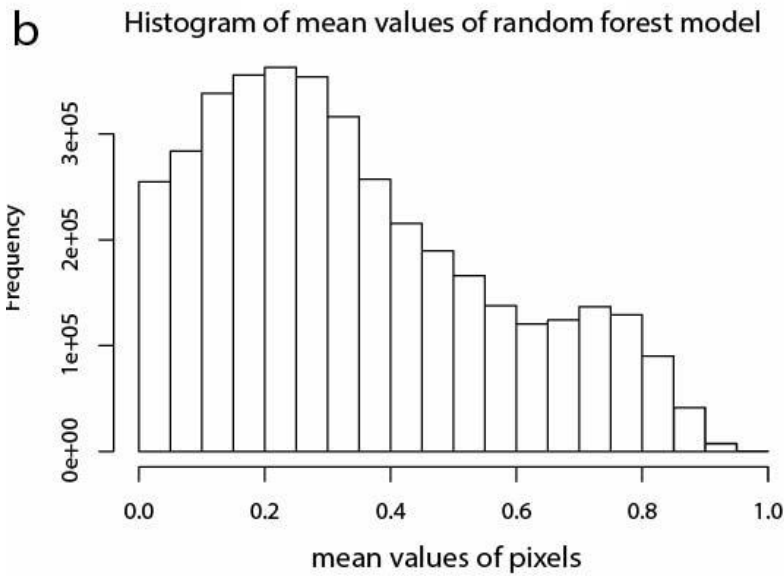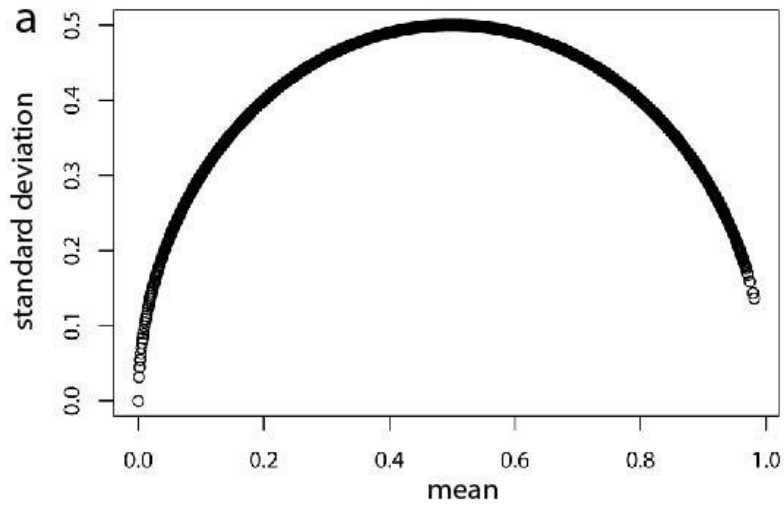Histogram of mean values of random forest model

FIGURE S4: Mean values of each pixel of the random forest model: (a) standard deviation plotted against mean, and (b) histogram of mean pixel values.
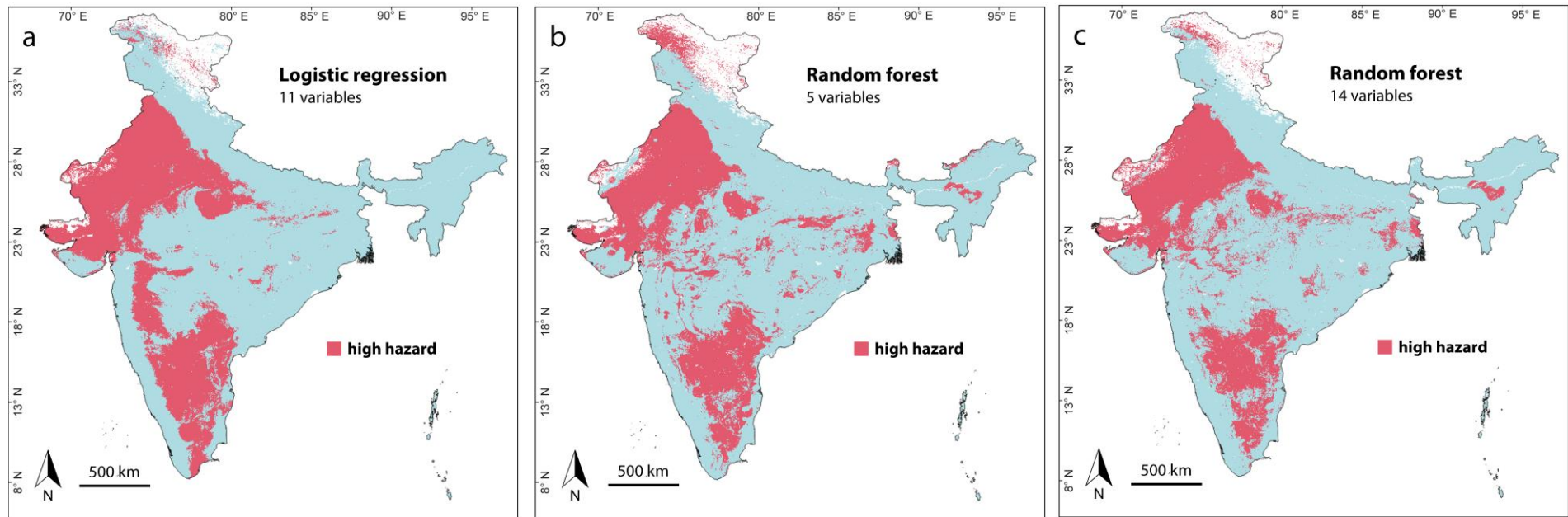
FIGURE S5. Comparison of high-hazard areas according the respective cutoff values calculated for each model: (a) logistic regression with the 11 independent variables listed in Table 4, (b) random forest with the five most frequent variables as listed in Table 4, and (c) random forest model with 14 variables as shown in Figure 2a.