

## **Supplementary Materials**

# **Geogenic manganese and iron in groundwater of Southeast Asia and Bangladesh – Machine learning spatial prediction modeling and comparison with arsenic**

**Authors:** Joel Podgorski<sup>1\*</sup>, Dahyann Araya<sup>1</sup>, Michael Berg<sup>1</sup>

<sup>1</sup> Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland

\* corresponding author, e-mail: joel.podgorski@eawag.ch

prepared March 2022

### **Contents:**

Table S1: Numbers of Mn, Fe and physicochemical parameters

Table S2: Spatially continuous predictor variables

Table S3: Performance statistics of random forest models with physicochemical variables

Figure S1: Kendall correlations between Mn, Fe and other measured parameters

Figure S2: PDPs of physicochemical predictor variables used in RF model for Mn

Figure S3: PDPs of physicochemical predictor variables used in RF model for Fe

Figure S4: PDPs of spatially continuous predictor variables in Mn model

Figure S5: PDPs of spatially continuous predictor variables in Fe model

Figure S6: Maps of the more important variables used in Mn/Fe spatial prediction models

Figure S7: Probability map of Mn > 400 µg/L for Southeast Asia and Bangladesh

Figure S8: Probability map of Fe > 0.3 mg/L for Southeast Asia and Bangladesh

Figure S9: Probability map of As > 10 µg/L for Southeast Asia and Bangladesh

Figure S10: Maps of coefficient of variation of Mn/Fe RF and GBM models

References in Supplementary Materials

**Table S1:** Number of manganese (Mn) and iron (Fe) measurements and other physicochemical parameters compiled in Southeast Asia and their data sources, sorted by area.

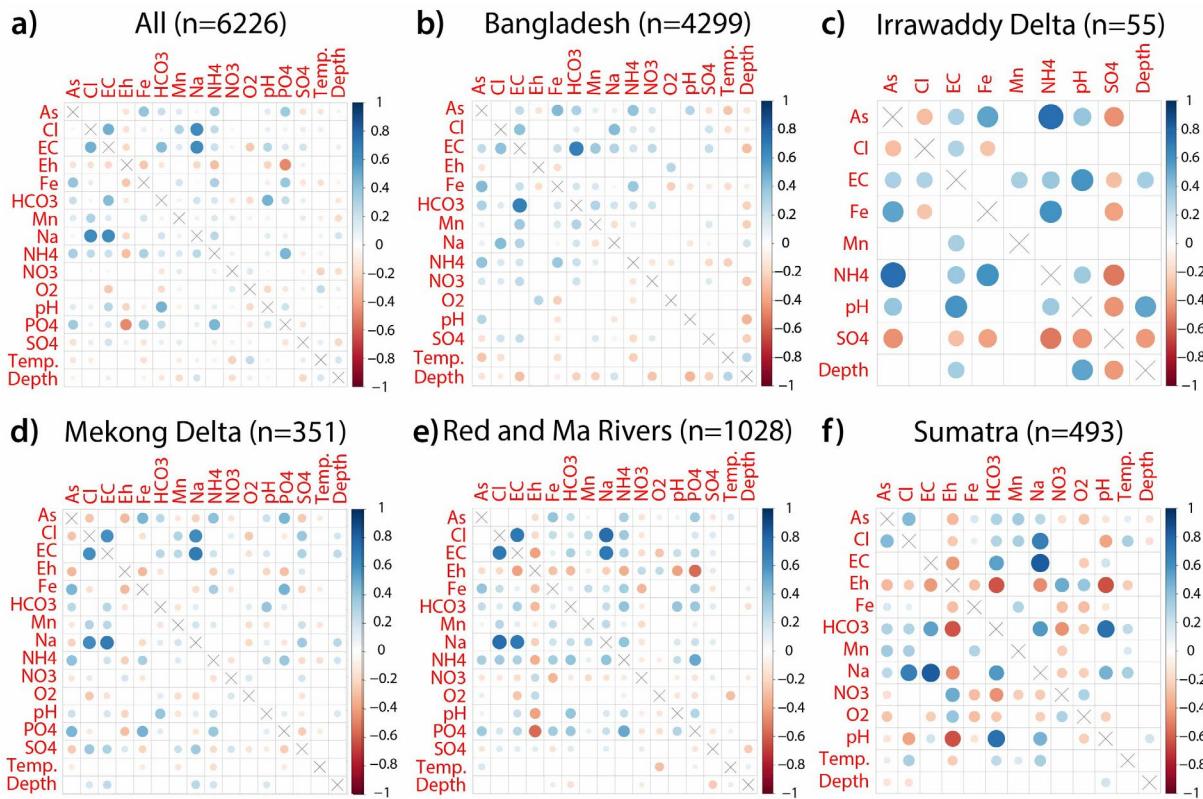
Parameter	Bangladesh	Mekong Delta (Cambodia, Vietnam)	Myanmar	Red and Ma River Deltas (Vietnam))	Sumatra (Indonesia)	TOTAL
Mn	4213	351	55	1028	475	6122
Fe	4209	351	55	1028	464	6107
As	4299	350	55	1028	489	6221
Cl	466	350	55	557	368	1796
EC	355	350	55	978	387	2125
Eh	269	341	-	974	102	1686
HCO <sub>3</sub>	468	351	-	978	102	1899
Na	4209	351	-	512	101	5173
NH <sub>4</sub>	384	351	55	982	203	1975
NO <sub>3</sub>	468	350	-	511	332	1661
O <sub>2</sub>	268	308	-	808	102	1486
pH	256	350	55	976	387	2024
PO <sub>4</sub>	84	351	-	512	-	947
SO <sub>4</sub>	3998	351	55	558	406	5368
Temp.	271	350	-	508	387	1516
Well depth	4197	325	55	547	394	5518
Source	(BGS and DPHE, 2001; Hoque et al., 2014)	(Buschmann et al., 2008)	Van Geen et al., 2014)	(Berg et al., 2001; Buschmann et al., 2008; Winkel et al., 2011)	(Marohn et al., 2012; Winkel et al., 2008)	

**Table S2:** Spatially continuous predictor variables (n=57) with source and resolution (Res.). The variables contain continuous values, unless indicated as being categorical (cat.). Topsoil refers to 0 cm depth and subsoil to 200 cm depth. (At the equator, 30" is approximately equal to 1 km.)

Parameter	Res.	Parameter	Res.
Climate		Soil (cont.)	
Actual evapotranspiration (AET) (Trabucco and Zomer, 2010)	30"	Organic carbon volume, subsoil (Hengl et al., 2017)	7.5"
Aridity (PET (Trabucco and Zomer, 2009) / precipitation (Fick Stephen and Hijmans Robert, 2017))	30"	pH, subsoil (measured in water) (Hengl et al., 2017)	7.5"
Potential evapotranspiration (PET) (Trabucco and Zomer, 2009)	30"	Regosols probability (Hengl et al., 2017)	7.5"
Precipitation (Fick Stephen and Hijmans Robert, 2017)	30"	Sand (0.05-2 mm), weight %, topsoil and subsoil (Hengl et al., 2017)	7.5"
Priestley-Taylor alpha coefficient (AET/PET) (Trabucco and Zomer, 2010)	30"	Silt (0.0002-0.05 mm), weight %, topsoil and subsoil (Hengl et al., 2017)	7.5"
Temperature (Fick Stephen and Hijmans Robert, 2017)	30"	Solonchaks probability (Hengl et al., 2017)	7.5"
Other		USDA soil texture classes (cat.), topsoil and subsoil (Hengl, 2018d)	7.5"
Irrigation amounts (Vörösmarty et al., 2005)	30'	Water content (volumetric %) for 33kPa and 1500kPa, topsoil and subsoil (Hengl and Gupta, 2019)	7.5"
Landcover (Friedl et al., 2010)	15"	Water capacity until wilting points (volumetric %), subsoil (Hengl et al., 2017)	7.5"
Lithology classes (cat.) (Hengl, 2018b)	7.5"		
Water table depth (Fan et al., 2013)	30"	Topography	
Soil		Compound topographic index (Amatulli et al., 2019)	7.5"
Acrisols probability	7.5"	Convergence index (Amatulli et al., 2019)	7.5"
Alisols probability (Hengl et al., 2017)	7.5"	Downslope curvature (Hengl, 2018a)	7.5"
Arenosols probability (Hengl et al., 2017)	7.5"	Elevation (Hengl, 2018a)	7.5"
Calcisols probability (Hengl et al., 2017)	7.5"	Flow accumulation (Lehner et al., 2006)	30"
Cation exchange capacity (Hengl et al., 2017)	7.5"	Geomorphometric classes (cat.) (Amatulli et al., 2019)	7.5"
Clay (<0.0002 mm), weight %, topsoil and subsoil (Hengl et al., 2017)	7.5"	Maximum multiscale deviation (Amatulli et al., 2019)	7.5"
Coarse fragments (>2 mm), vol. %, topsoil and subsoil (Hengl et al., 2017)	7.5"	Maximum multiscale roughness (Amatulli et al., 2019)	7.5"
FAO soil classes (cat.) (Hengl et al., 2017)	7.5"	Profile curvature (Amatulli et al., 2019)	7.5"
Fine earth bulk density, topsoil and subsoil (Hengl, 2018c)	7.5"	Roughness (Amatulli et al., 2019)	7.5"
Fluvisols probability (Hengl et al., 2017)	7.5"	Scale of the maximum multiscale roughness (Amatulli et al., 2019)	7.5"
Gleysols probability (Hengl et al., 2017)	7.5"	Tangential curvature (Amatulli et al., 2019)	7.5"
Hydrologic soil groups (cat.) (Ross et al., 2018)	7.5"	Terrain ruggedness index (Amatulli et al., 2019)	7.5"
Organic carbon density, subsoil (Hengl et al., 2017)	7.5"	Valley bottom flatness (Hengl, 2018a)	7.5"
Organic carbon stock loss(kg/m <sup>2</sup> ), topsoil, 2001-2015 (Wheeler and Hengl, 2018)	7.5"	Vector ruggedness measure (Amatulli et al., 2019)	7.5"

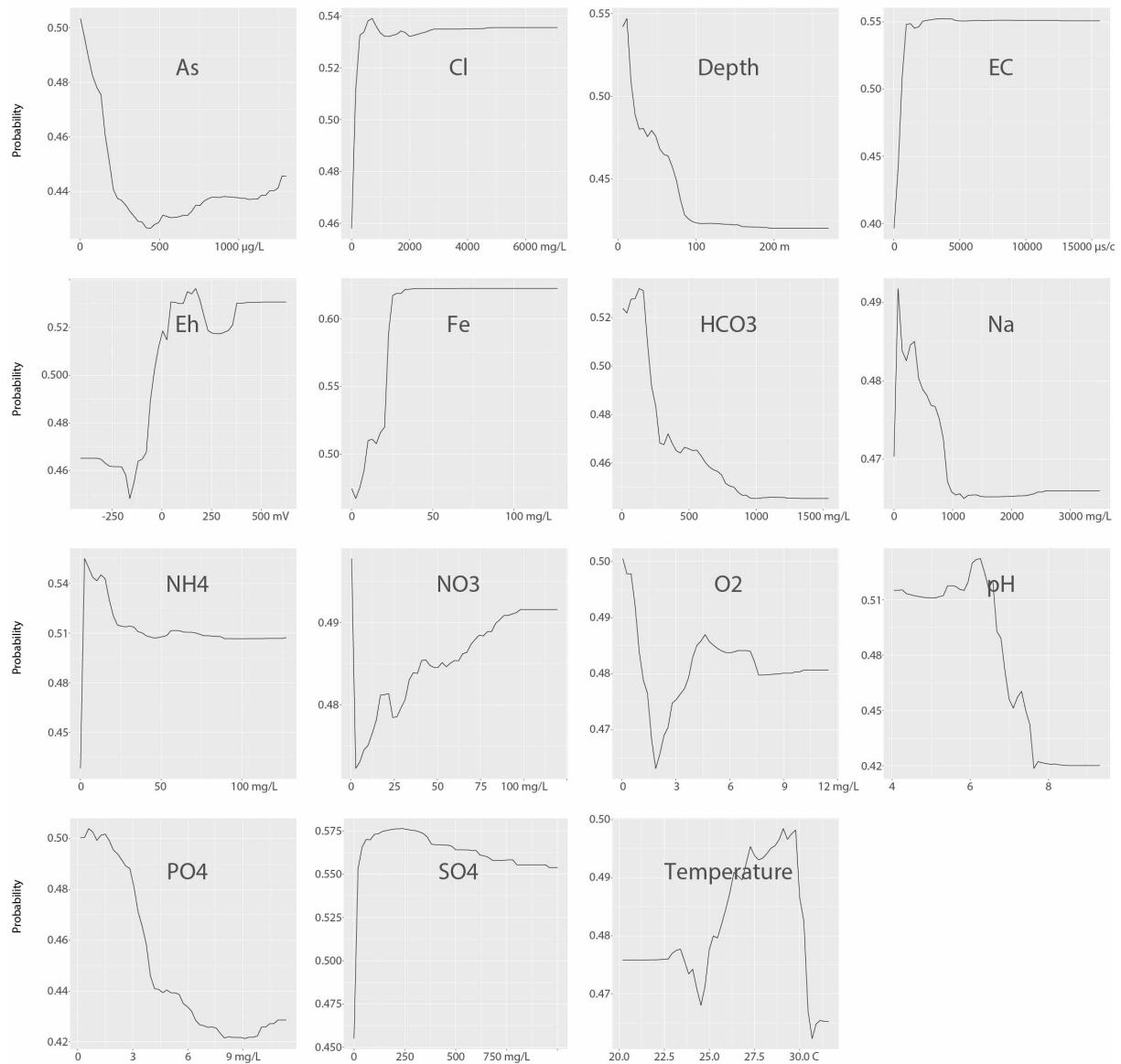
**Table S3:** Performance statistics of random forest models of Mn > 400 µg/L and Fe > 0.3 mg/L using physicochemical variables as applied to the test datasets.

	Mn (RF)	Fe (RF)
<b>Prob. cutoff</b>	0.48 ± 0.02	0.48 ± 0.06
<b>Balanced accuracy</b>	0.69 ± 0.04	0.88 ± 0.03
<b>Kappa</b>	0.39 ± 0.08	0.75 ± 0.05
<b>AUC</b>	0.78 ± 0.04	0.95 ± 0.02



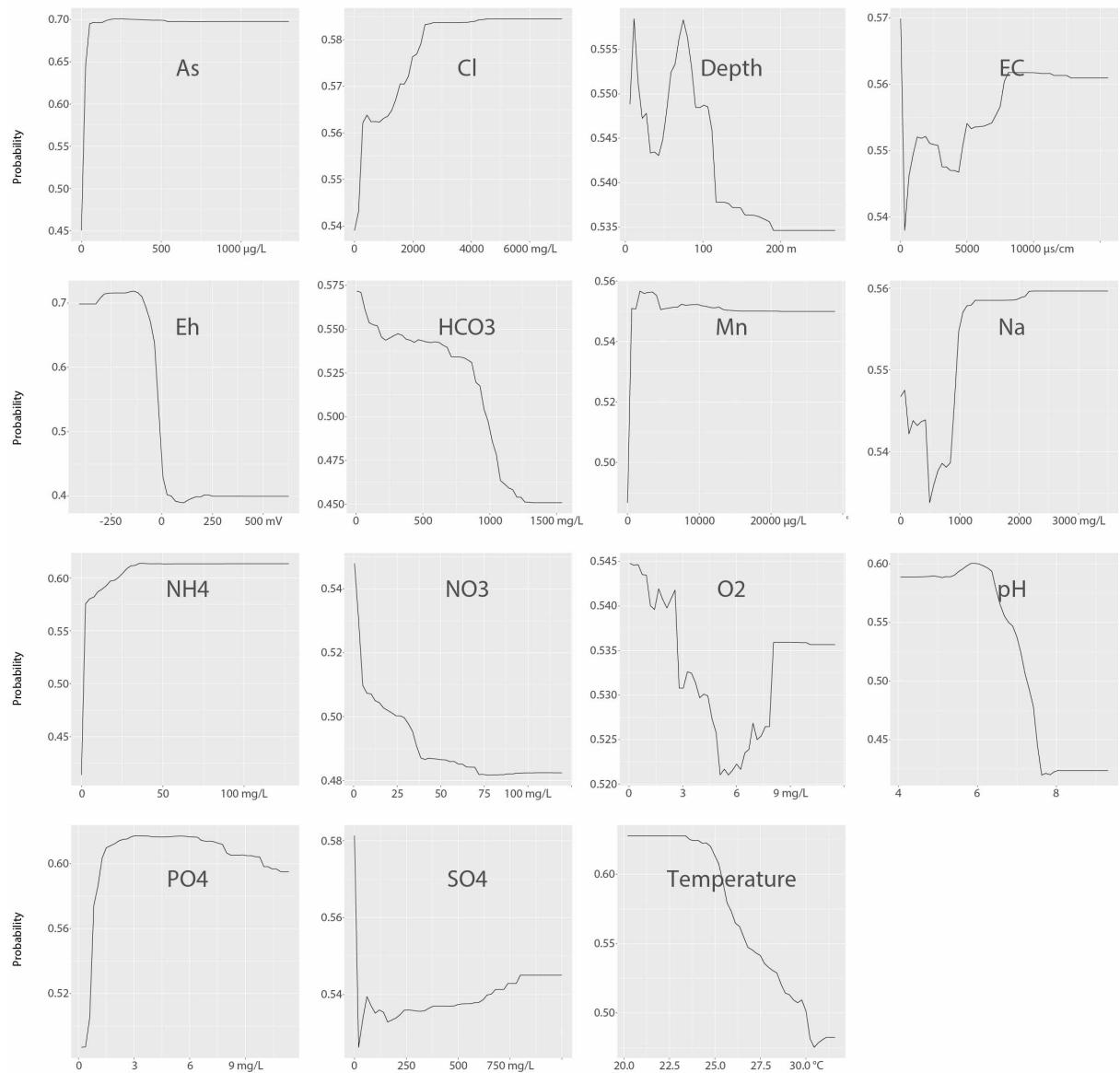
**Figure S1:** Kendall correlations between manganese, iron and 14 other measured parameters at the 0.05 significance level. Blank cells indicate the calculated correlation did not meet the 0.05 significance level. Correlations could not be calculated of all the parameters for some locations due to insufficient data.

### Mn (RF - physicochemical predictors)



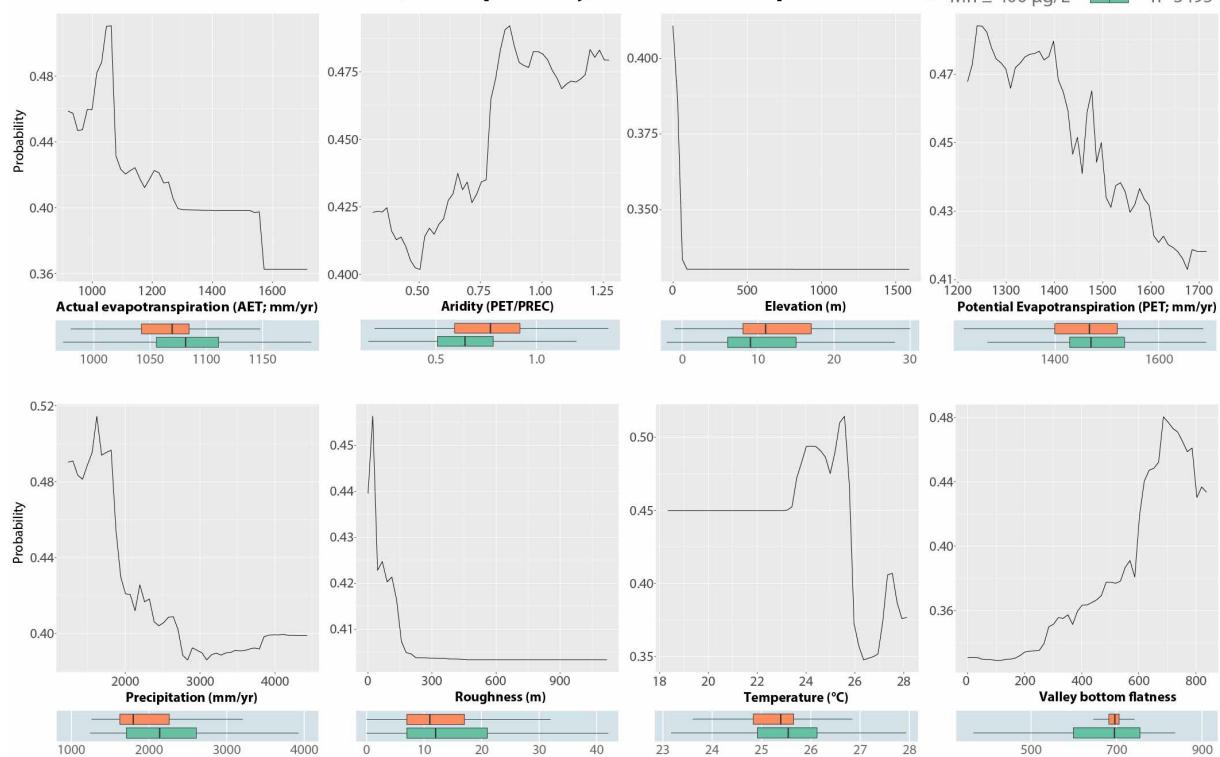
**Figure S2:** Partial dependence plots (PDP) of the 15 physicochemical predictor variables used in the random forest model for Mn in groundwater.

### Fe (RF - physicochemical predictors)



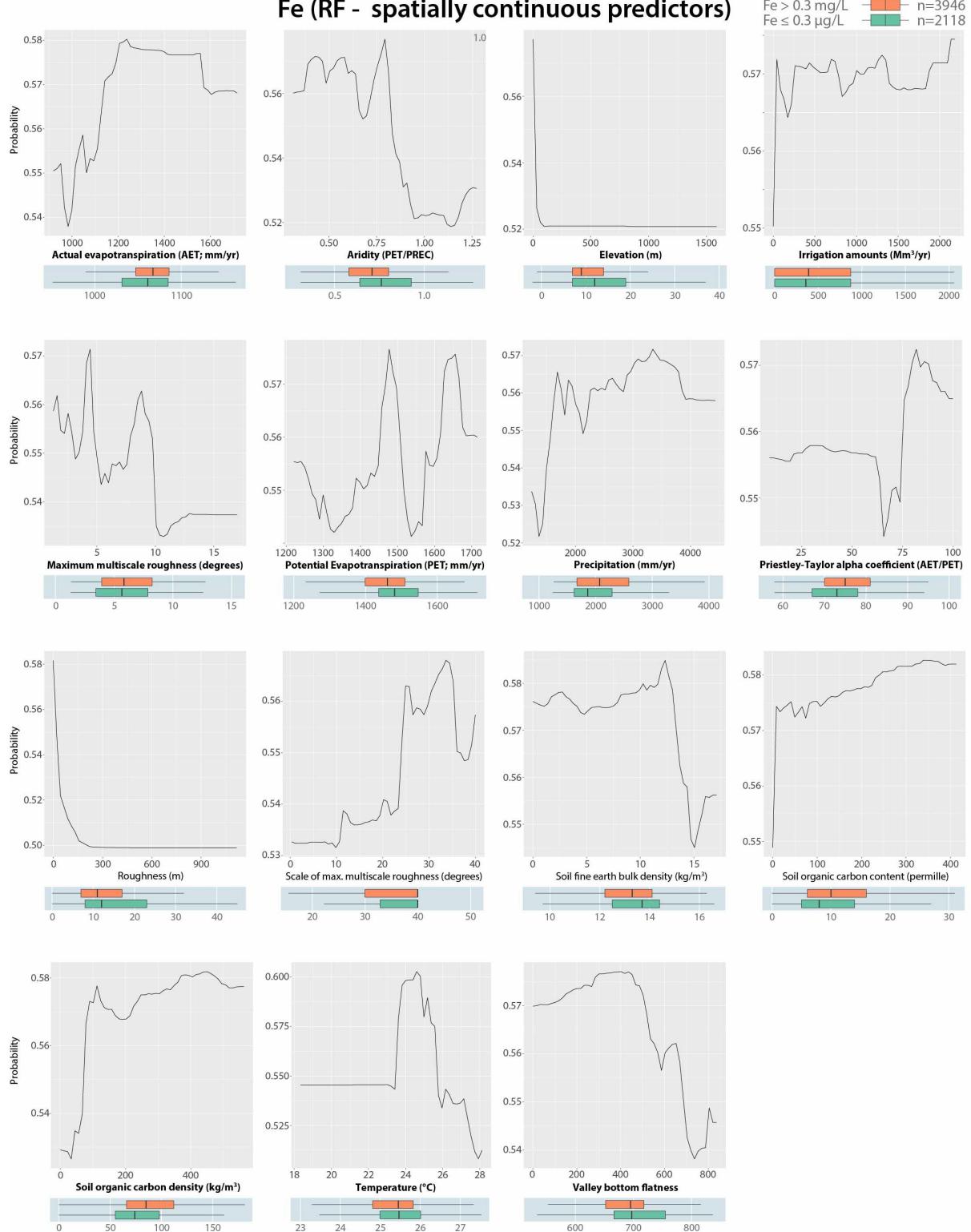
**Figure S3:** Partial dependence plots (PDP) of the 15 physicochemical predictor variables used in the random forest model for Fe in groundwater.

### Mn (RF - spatially continuous predictors)

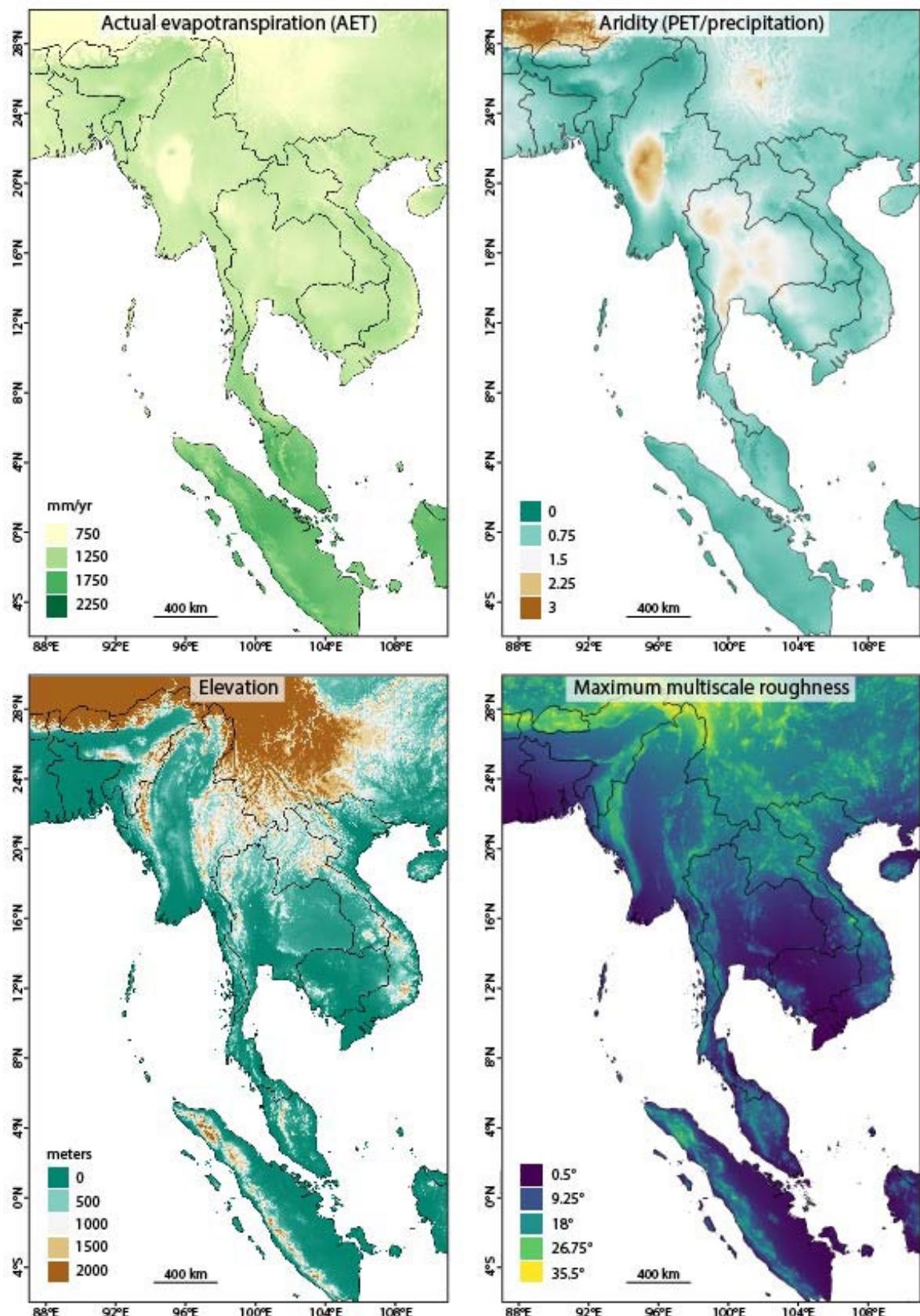


**Figure S4:** Partial dependence plots (PDP) of the eight spatially continuous predictor variables used in the random forest model for Mn in groundwater. Below each PDP are boxplots of the distribution of predictor values associated with Mn measurements above and below 400 µg/L. The thicker gray bars mark the median, the boxes indicate the 0.25 and 0.75 percentiles and the whiskers indicate the lowest and highest values within 1.5 times the interquartile range.

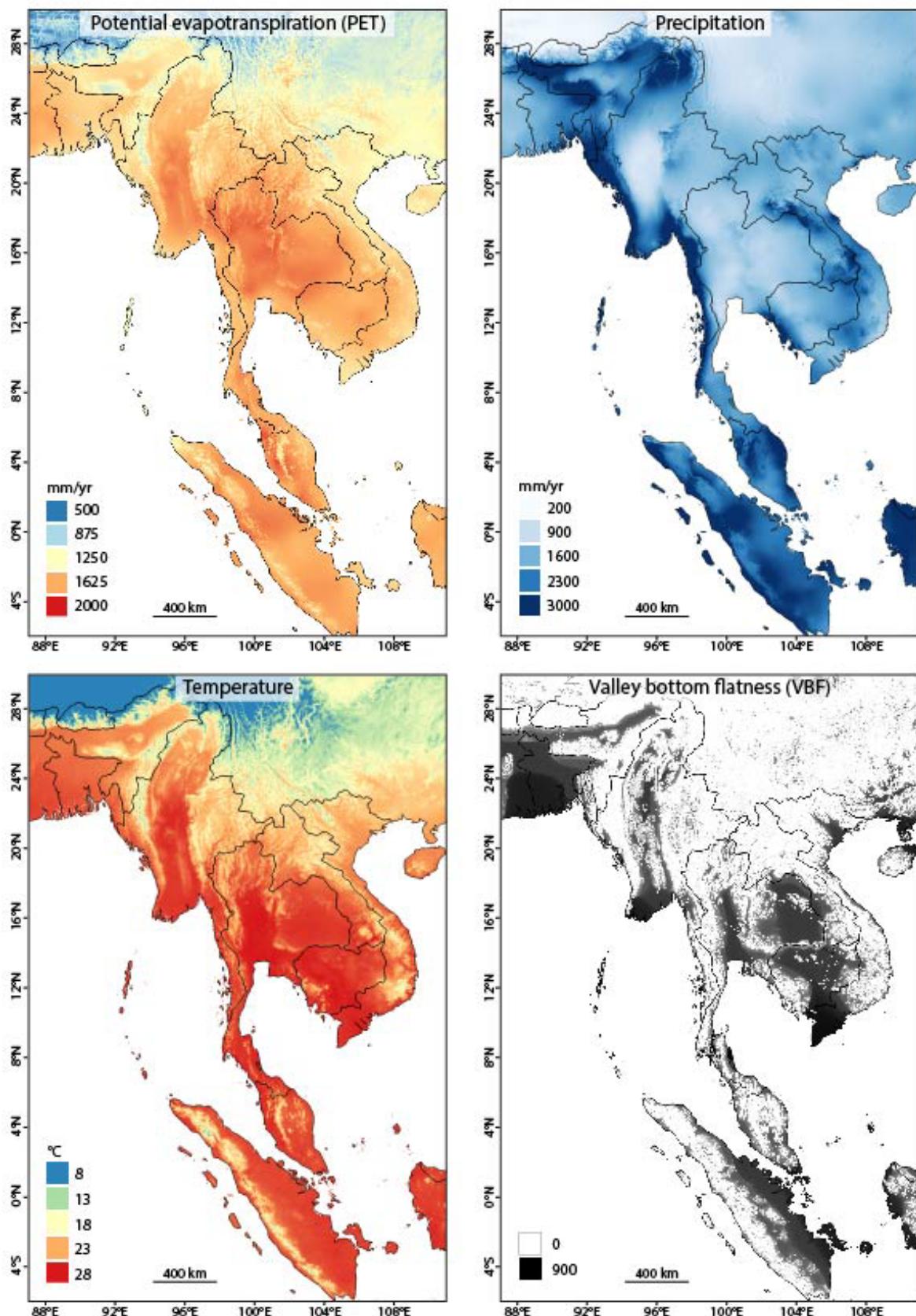
### Fe (RF - spatially continuous predictors)



**Figure S5:** Partial dependence plots (PDP) of the 15 spatially continuous predictor variables used in the random forest model for Fe in groundwater. Below each PDP are boxplots of the distribution of predictor values associated with Fe measurements above and below 0.3 mg/L. The thicker gray bars mark the median, the boxes indicate the 0.25 and 0.75 percentiles and the whiskers indicate the lowest and highest values within 1.5 times the interquartile range.



**Figure S6:** Maps of eight of the more important predictor variables used in the RF spatial prediction models of groundwater Mn and Fe (Figs. 5, 6).



**Figure S6 (cont.)**

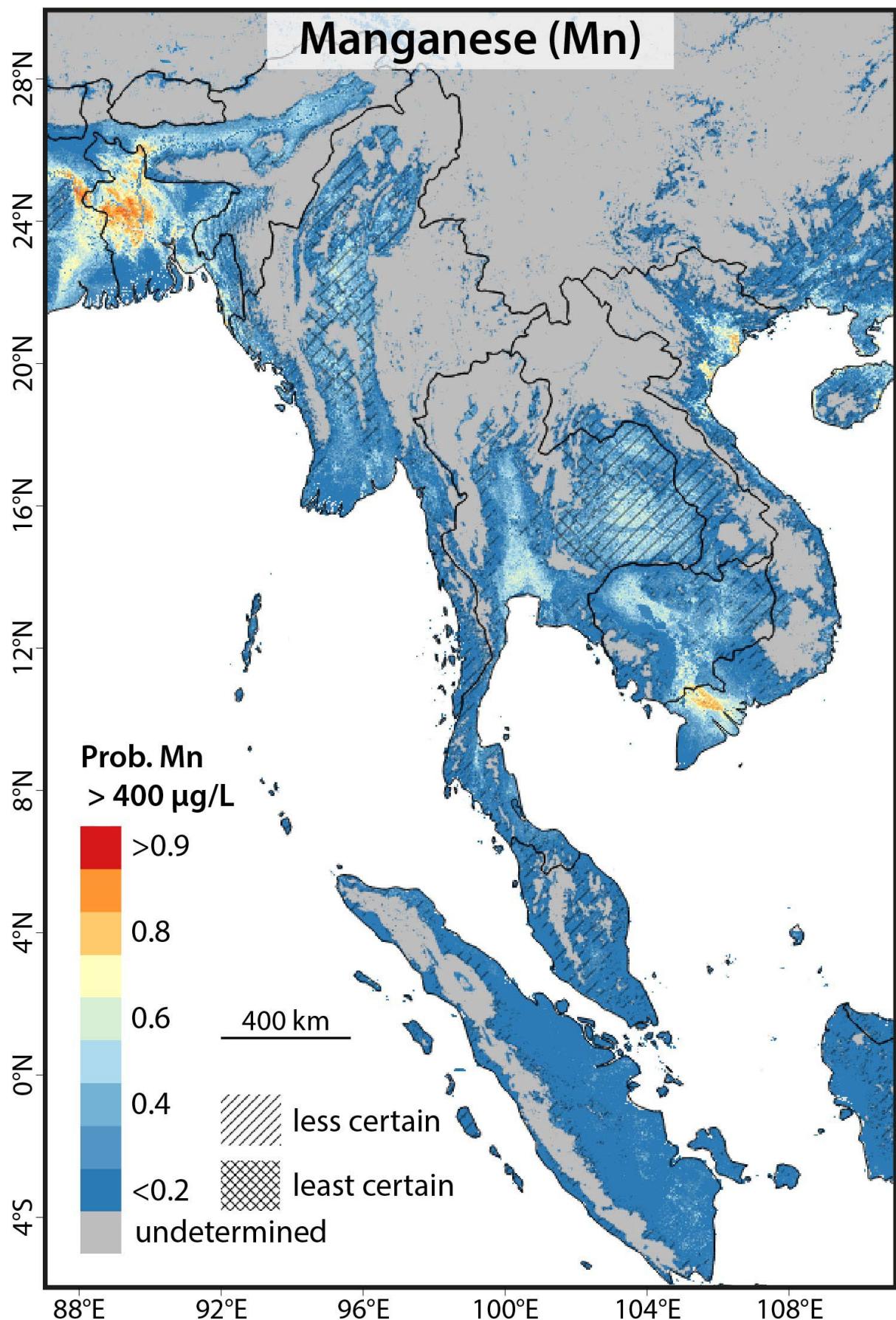
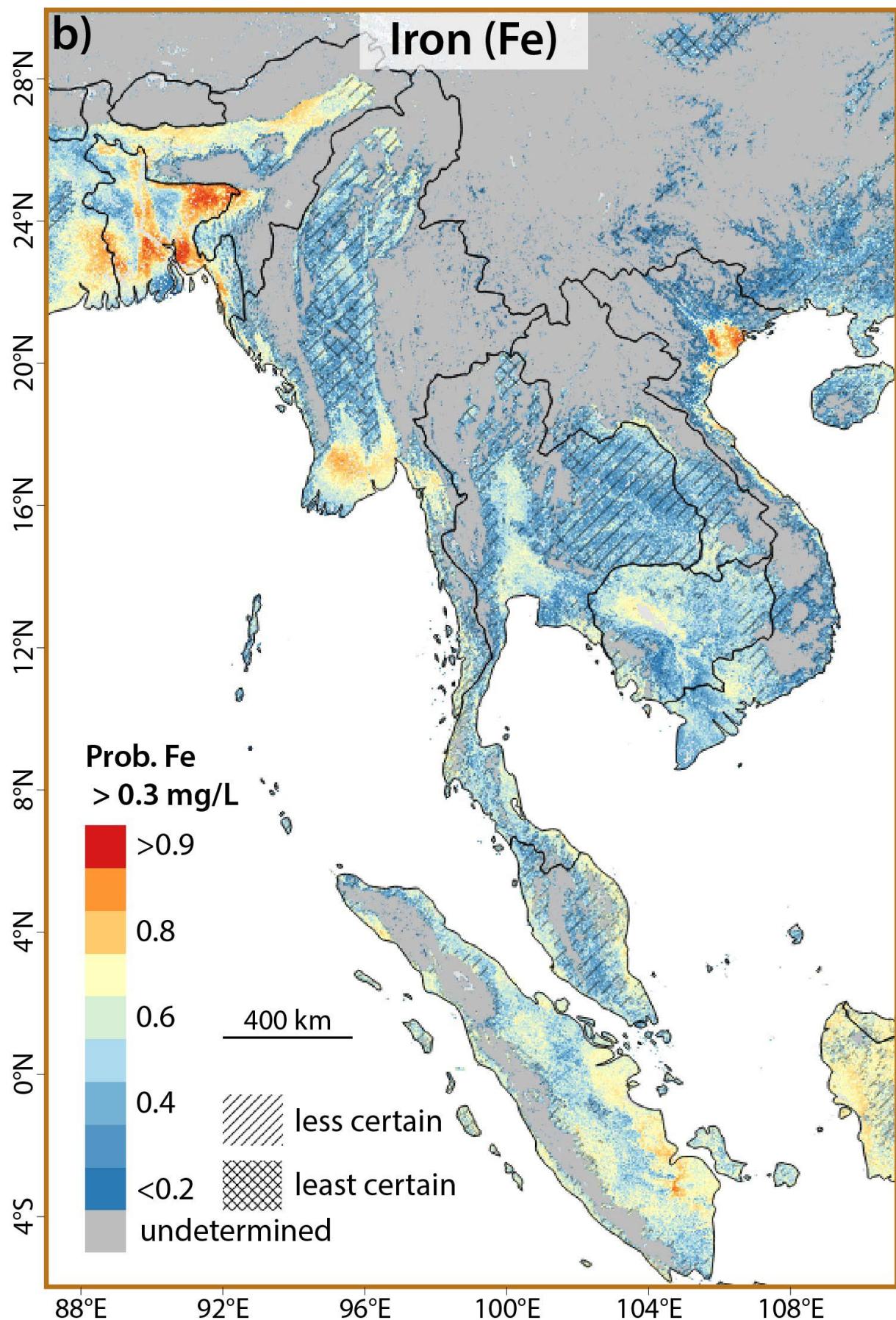
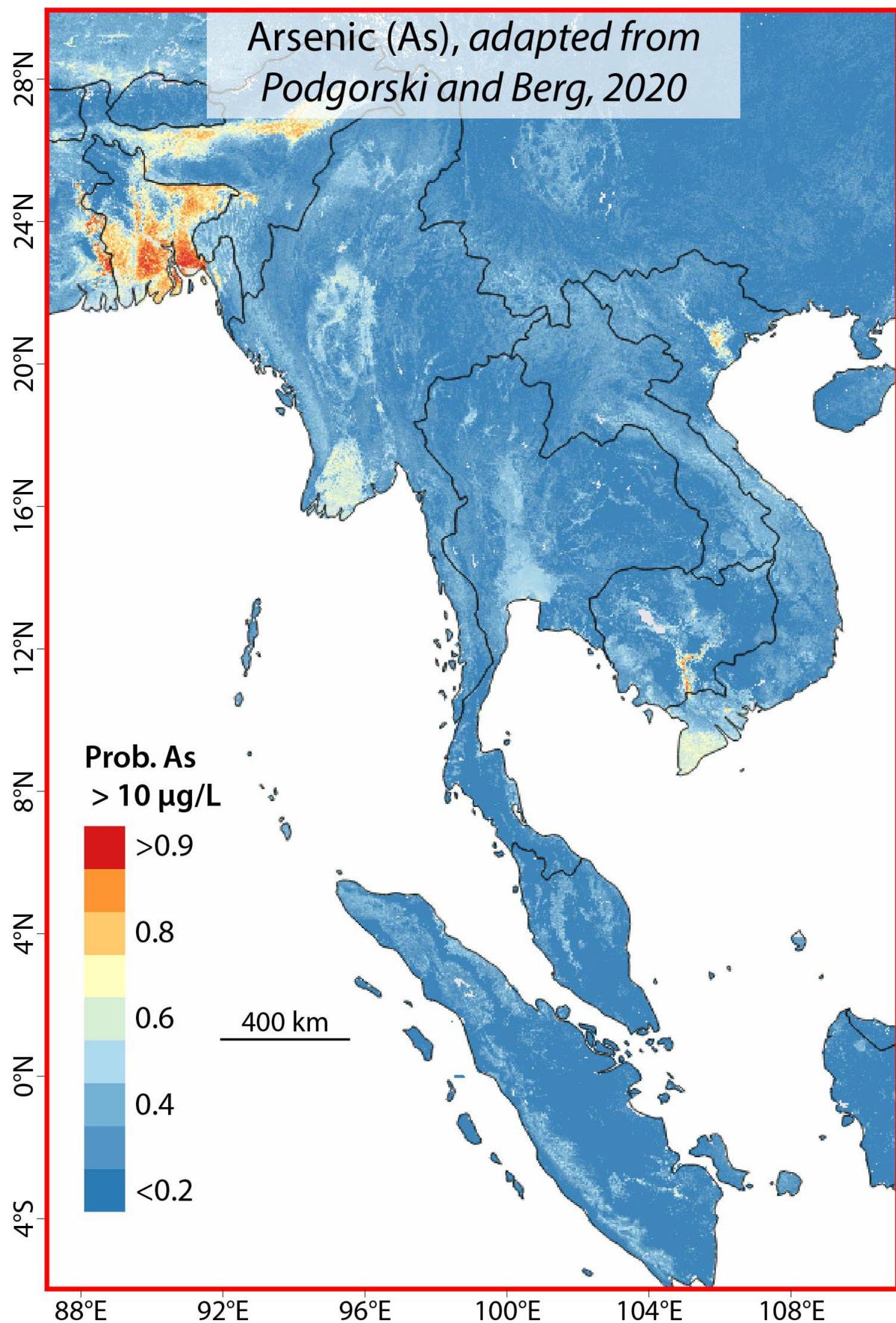


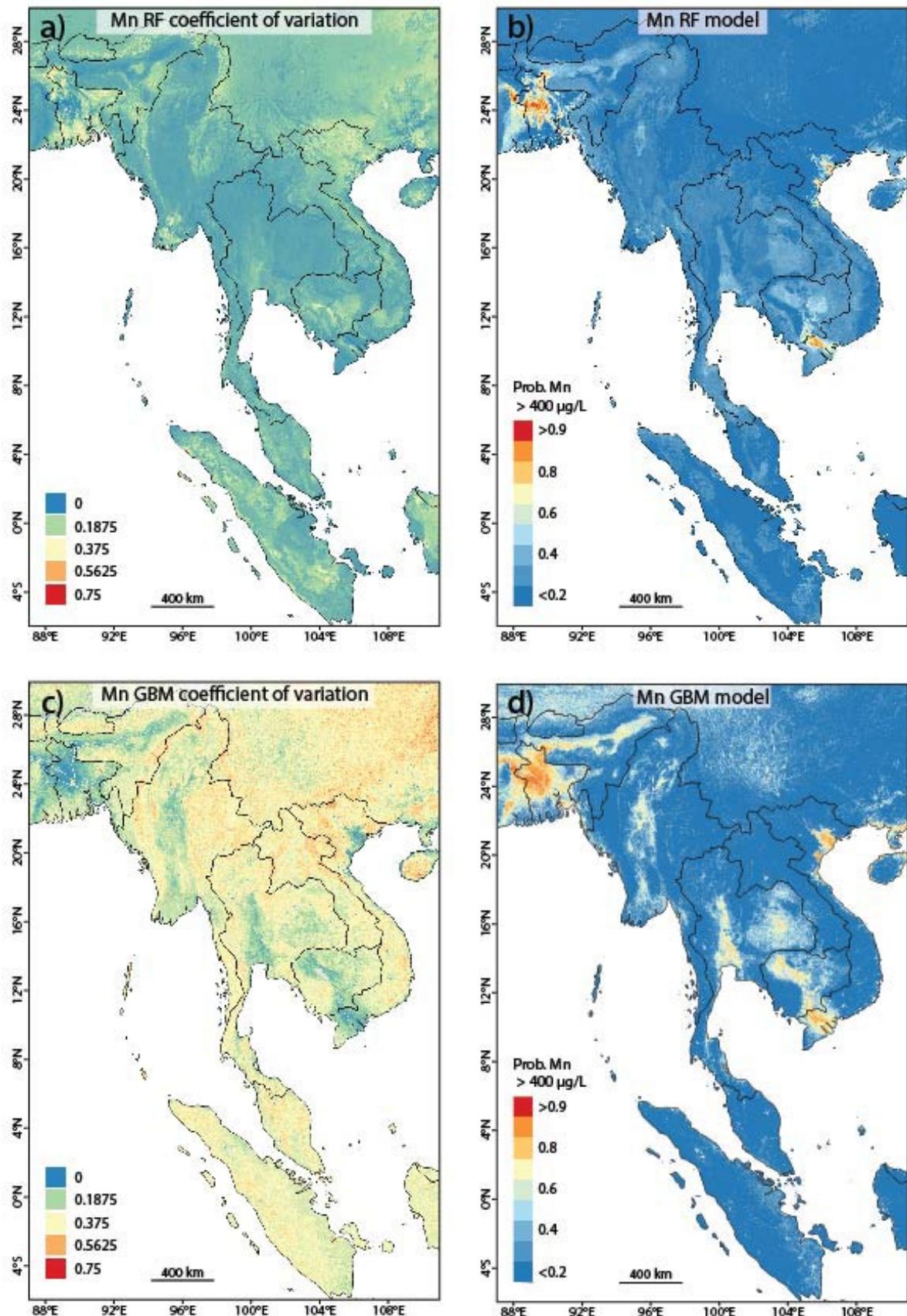
Figure S7: Probability map of Mn  $> 400 \mu\text{g/L}$  for Southeast Asia and Bangladesh.



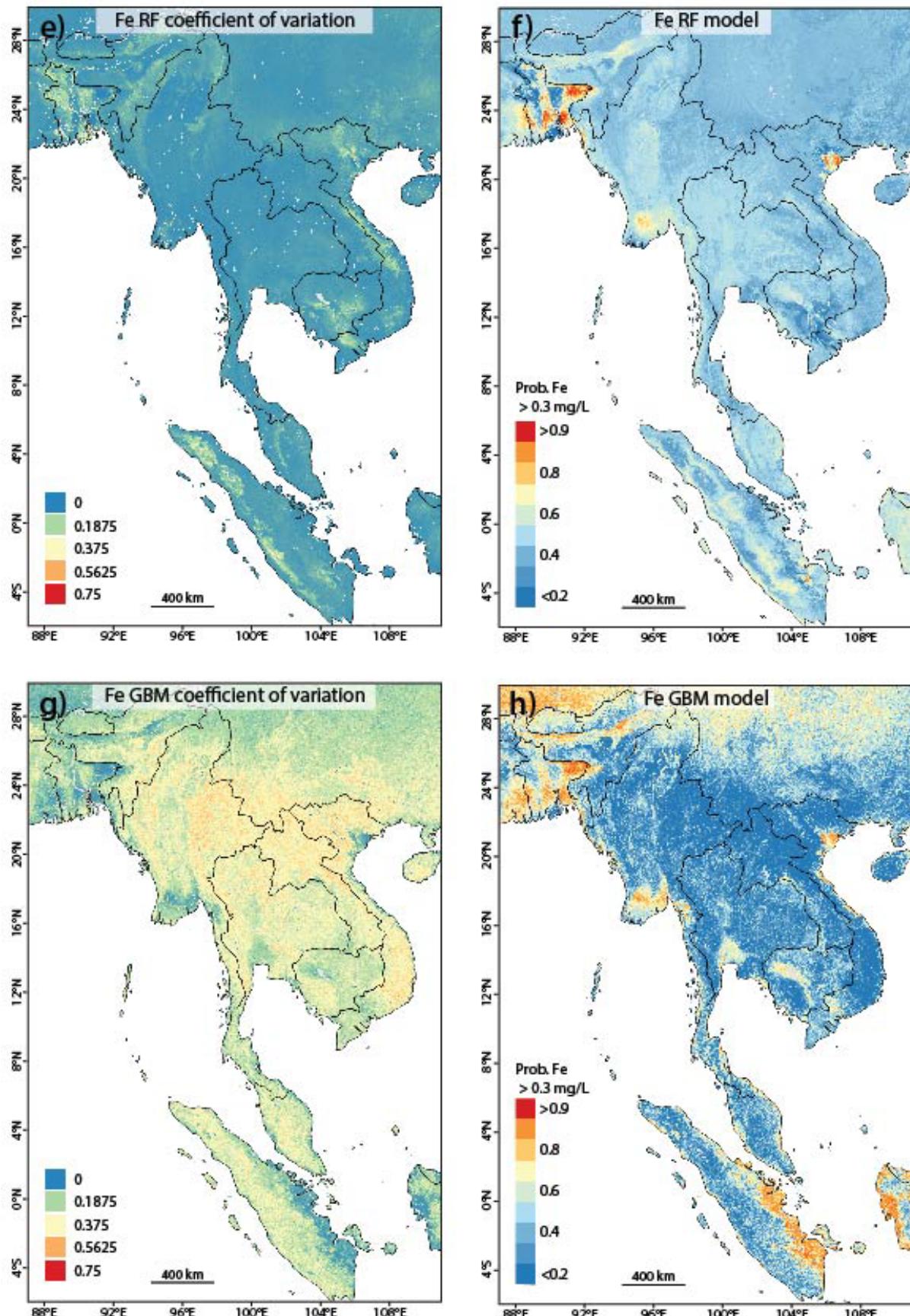
**Figure S8:** Probability map of Fe  $> 0.3 \text{ mg/L}$  for Southeast Asia and Bangladesh.



**Figure S9:** Probability map of As  $> 10 \mu\text{g/L}$  for Southeast Asia and Bangladesh, adapted from Podgorski and Berg, 2020.



**Figure S10:** Coefficient of variation (quotient of standard deviation and mean) of maps created from 100-fold cross validation along with the associated models of  $Mn > 400 \mu\text{g/L}$  for (a, b) RF and (c, d) GBM as well as of  $Fe > 0.3 \text{ mg/L}$  for (e, f) RF and (g, h) GBM.



**Figure S10 (cont.)**

## References in Supplementary Materials

- Amatulli G, McInerney D, Sethi T, Strobl P, Domisch S. Geomorpho90m-Global high-resolution geomorphometry layers: empirical evaluation and accuracy assessment. PeerJ Preprints, 2019.
- Berg M, Tran HC, Nguyen TC, Pham HV, Schertenleib R, Giger W. Arsenic contamination of groundwater and drinking water in Vietnam: a human health threat. Environmental Science & Technology 2001; 35: 2621-2626.
- BGS, DPHE. Arsenic contamination of groundwater in Bangladesh. In: Kinniburgh D, Smedley P, editors. British Geological Survey Technical Report. British Geological Survey, Keyworth, 2001.
- Buschmann J, Berg M, Stengel C, Winkel L, Sampson ML, Trang PTK, et al. Contamination of drinking water resources in the Mekong delta floodplains: Arsenic and other trace metals pose serious health risks to population. Environment International 2008; 34: 756-764.
- Fan Y, Li H, Miguez-Macho G. Global patterns of groundwater table depth. Science 2013; 339: 940-943.
- Fick Stephen E, Hijmans Robert JW. 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 2017.
- Friedl MA, Sulla-Menashe D, Tan B, Schneider A, Ramankutty N, Sibley A, et al. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote sensing of Environment 2010; 114: 168-182.
- Hengl T. Global DEM derivatives at 250 m, 1 km and 2 km based on the MERIT DEM (Version 1.0). <http://doi.org/10.5281/zenodo.1447210> 2018a.
- Hengl T. Global landform and lithology class at 250 m based on the USGS global ecosystem map (Version 1.0). <http://doi.org/10.5281/zenodo.1464846> 2018b.
- Hengl T. Soil bulk density (fine earth) 10 x kg / m-cubic at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. <http://doi.org/10.5281/zenodo.2525665> 2018c.
- Hengl T. Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m. <http://doi.org/10.5281/zenodo.2525817> 2018d.
- Hengl T, de Jesus JM, Heuvelink GB, Gonzalez MR, Kilibarda M, Blagotić A, et al. SoilGrids250m: Global gridded soil information based on machine learning. PLoS one 2017; 12: e0169748.
- Hengl T, Gupta S. Soil water content (volumetric %) for 33kPa and 1500kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution <http://doi.org/10.5281/zenodo.1447210> 2019.
- Hoque M, McArthur J, Sikdar P, Ball J, Molla T. Tracing recharge to aquifers beneath an Asian megacity with Cl/Br and stable isotopes: the example of Dhaka, Bangladesh. Hydrogeology journal 2014; 22: 1549-1560.
- Lehner B, Verdin K, Jarvis A. HydroSHEDS Technical Documentation. World Wildlife Fund US, Washington, DC., 2006, pp. Available at <http://hydrosheds.cr.usgs.gov>.
- Marohn C, Distel A, Tomlinson R, Noordwijk Mv, Cadisch G. Impacts of soil and groundwater salinization on tree crop performance in post-tsunami Aceh Barat, Indonesia. Natural Hazards and Earth System Sciences 2012; 12: 2879.
- Ross CW, Prihodko L, Anchang J, Kumar S, Ji W, Hanan NP. HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling. Scientific data 2018; 5: 180091.
- Trabucco A, Zomer R. Global soil water balance geospatial database. CGIAR Consortium for Spatial Information, Published online, available from the CGIAR-CSI GeoPortal at: <http://www.cgiar-csi.org> (last access: January 2013) 2010.
- Trabucco A, Zomer RJ. Global aridity index (global-aridity) and global potential evapo-transpiration (global-PET) geospatial database. CGIAR Consortium for Spatial Information 2009.
- Vörösmarty CJ, Léveque C, Revenga C, Bos R, Caudill C, Chilton J, et al. Fresh water. Millennium ecosystem assessment 2005; 1: 165-207.

- Wheeler I, Hengl T. Soil organic carbon stock (0–30 cm) in kg/m<sup>2</sup> time-series 2001–2015 based on the land cover changes. <http://doi.org/10.5281/zenodo.2529721> 2018.
- Winkel L, Berg M, Stengel C, Rosenberg T. Hydrogeological survey assessing arsenic and other groundwater contaminants in the lowlands of Sumatra, Indonesia. *Applied Geochemistry* 2008; 23: 3019-3028.
- Winkel LH, Trang PTK, Lan VM, Stengel C, Amini M, Ha NT, et al. Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century. *Proceedings of the National Academy of Sciences* 2011; 108: 1246-1251.