Figure 1 Flow diagram of the proposed methodology to predict the spatial variability of SOC across Mexico and CONUS. SOC datasets and SOC covariates are harmonized on a geographical information system and regression matrices were created. Then we compare different modeling approaches selecting the best algorithm (higher accuracy) given the available combination of data and covariates. We performed our 'best algorithm' under a variable reduction strategy (recursive feature elimination technique) and use the best predictors under a simulated annealing regression framework for generating SOC maps across 250m grids. These analyses were cross validated using the Random Forest prediction error as accuracy metric. We applied the same methodology to different periods of data availability. Residual analyses on independent datasets were performed for quantifying the different sources of uncertainty. The prediction limits are quantified using quantile regression forests, we also report the variance of all generated maps, we analyzed the spatial structure of residuals and predict their spatial trends (Ordinary Kriging) using both independent datasets and values from six different BD pedotransfer functions used to calculate SOC stocks.

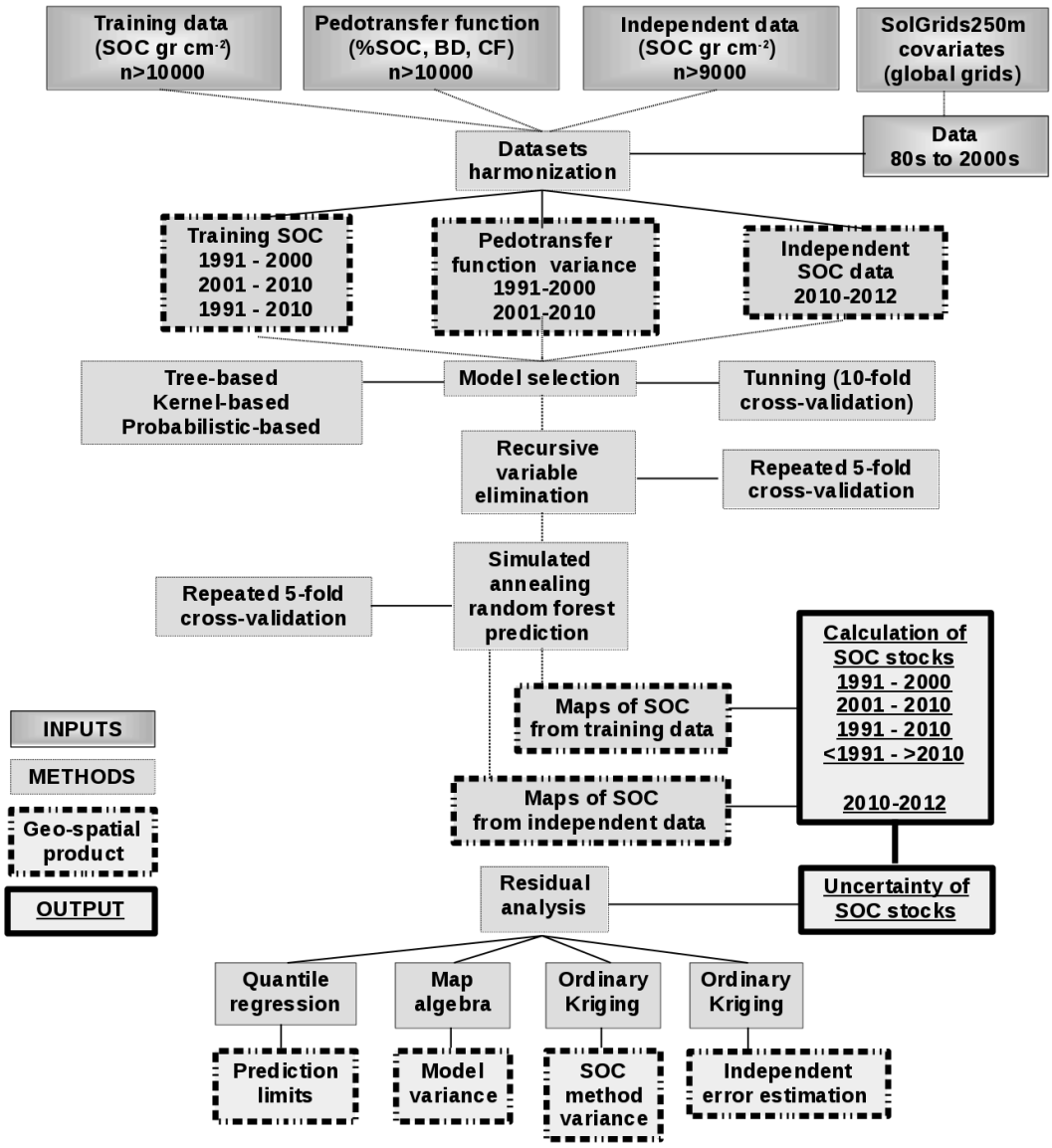


Figure 2 Spatial distribution of available datasets. We show the spatial distribution of available data for the period 1991-2010 (1991-2000 and 2001-2010) (A). A combined histogram of all datasets (combined CONUS and Mexico) shows that larger SOC values are mainly in the CONUS datasets while the higher density of values between 0 -1 is across Mexico (B). We found a moderate spatial structure of SOC available data that shows a large component of uncorrelated variation (nugget) probably due to the combination of different SOC sampling efforts.

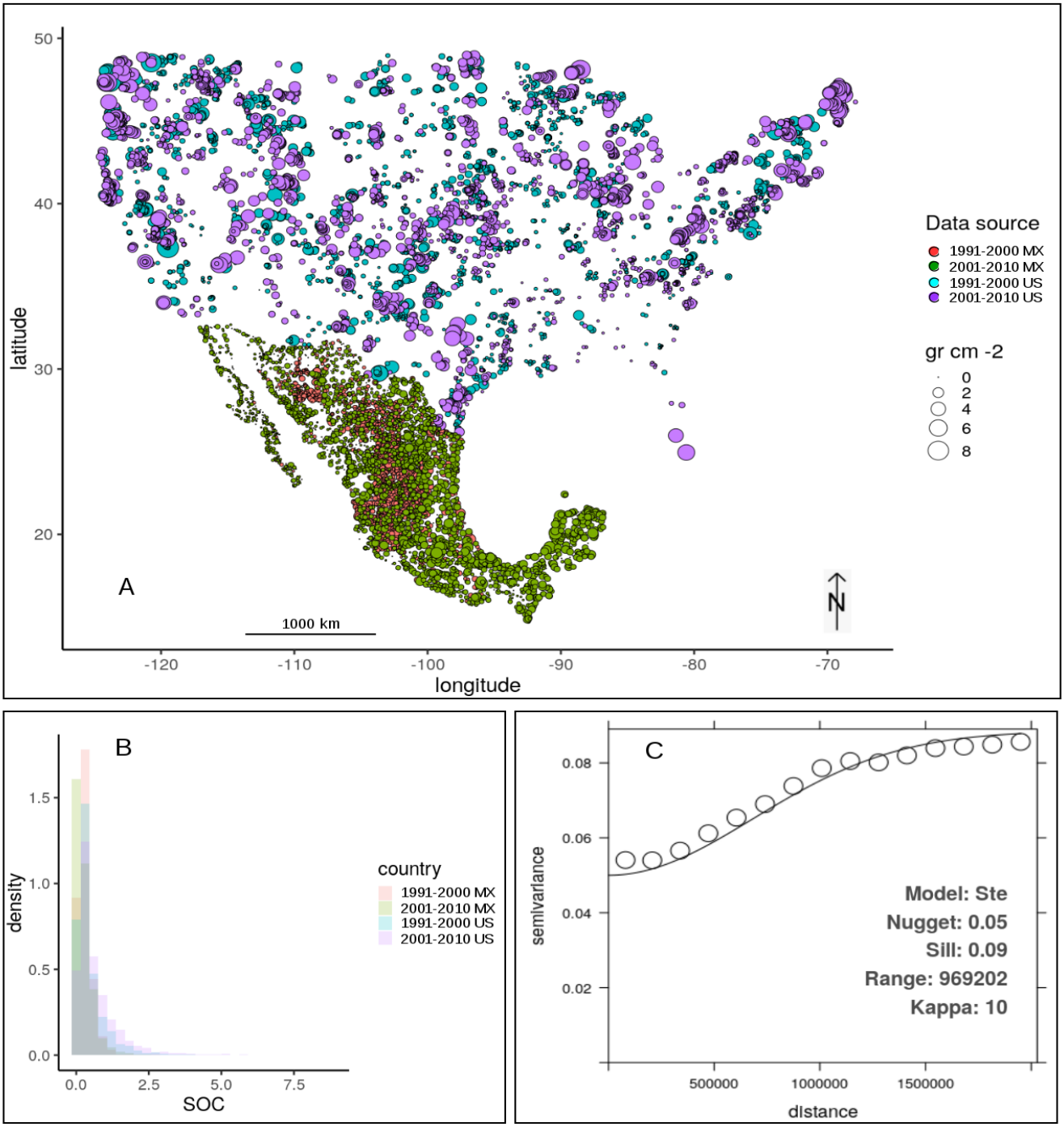


Figure 3 Visualization example across northern California and south Oregon in CONUS for the best four informative prediction factors for SOC. We show a MODIS longterm average of land surface temperature (degrees) (A), the long term precipitation mean value from the WorldClim initiative (B), a MODIS precipitable water vapor estimate (C) and the digital elevation model from the SRTM mission (D). The later is shown because it is the source for the calculation of the different terrain parameters that were best ranked predictors for SOC (e.g., wetness index, valley bottom flatness index).

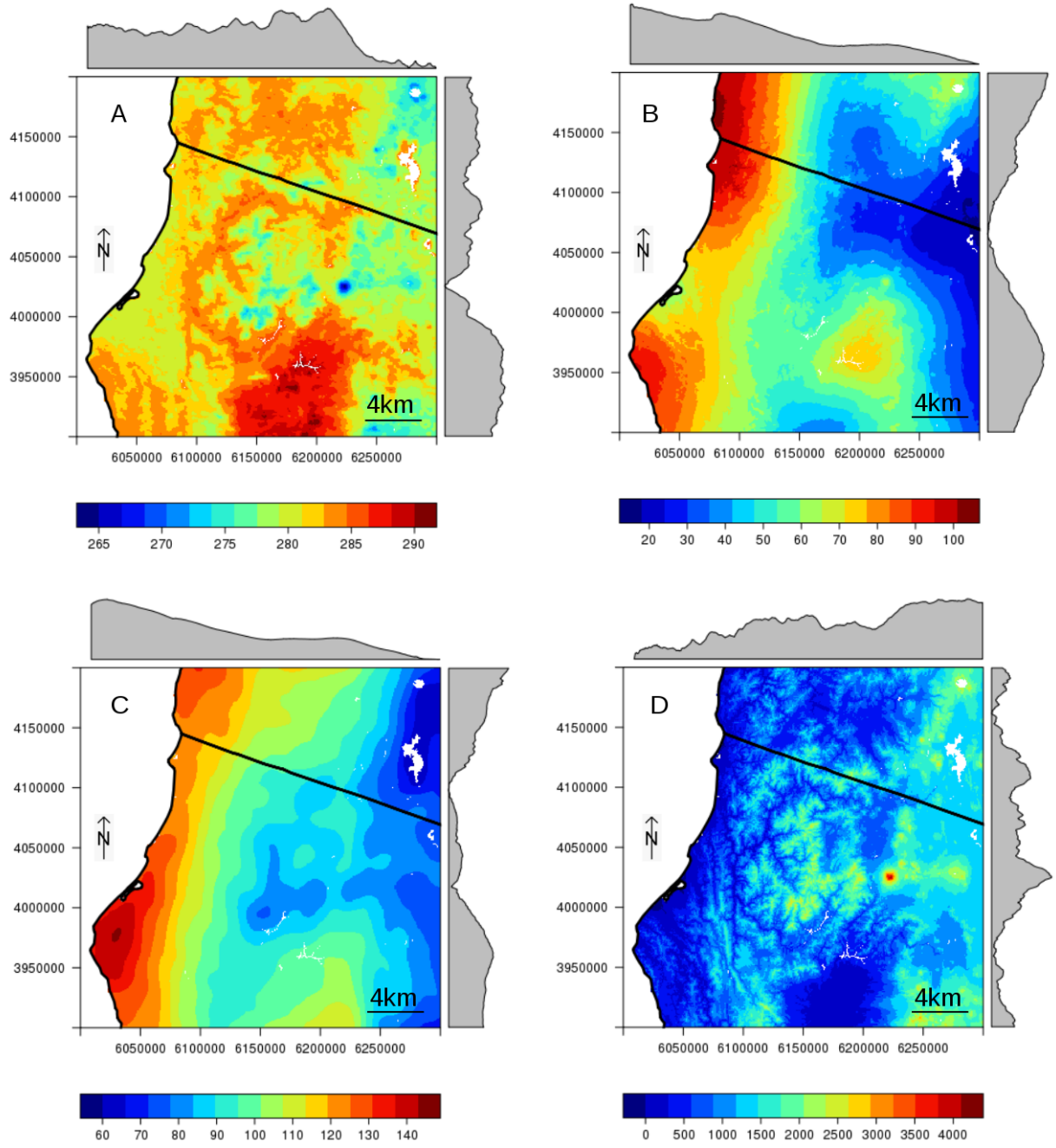


Figure 4 Spatial analysis applied to residuals of SOC models. The variogram of residuals against independent datastes (A). The residual variance from the different pedotransfer functions used to calculate SOC stocks (B). The combined (independent models and pedotransfer functions) residuals for the periods 1991-2000 (C) and 2001-2010 (D). The numbers in the circles indicate the available pars of points at a given distance. We also show the variogram parameters as insets on each plot. Sph = spherical model, Ste = Matern family, M. Stein's parameterization.

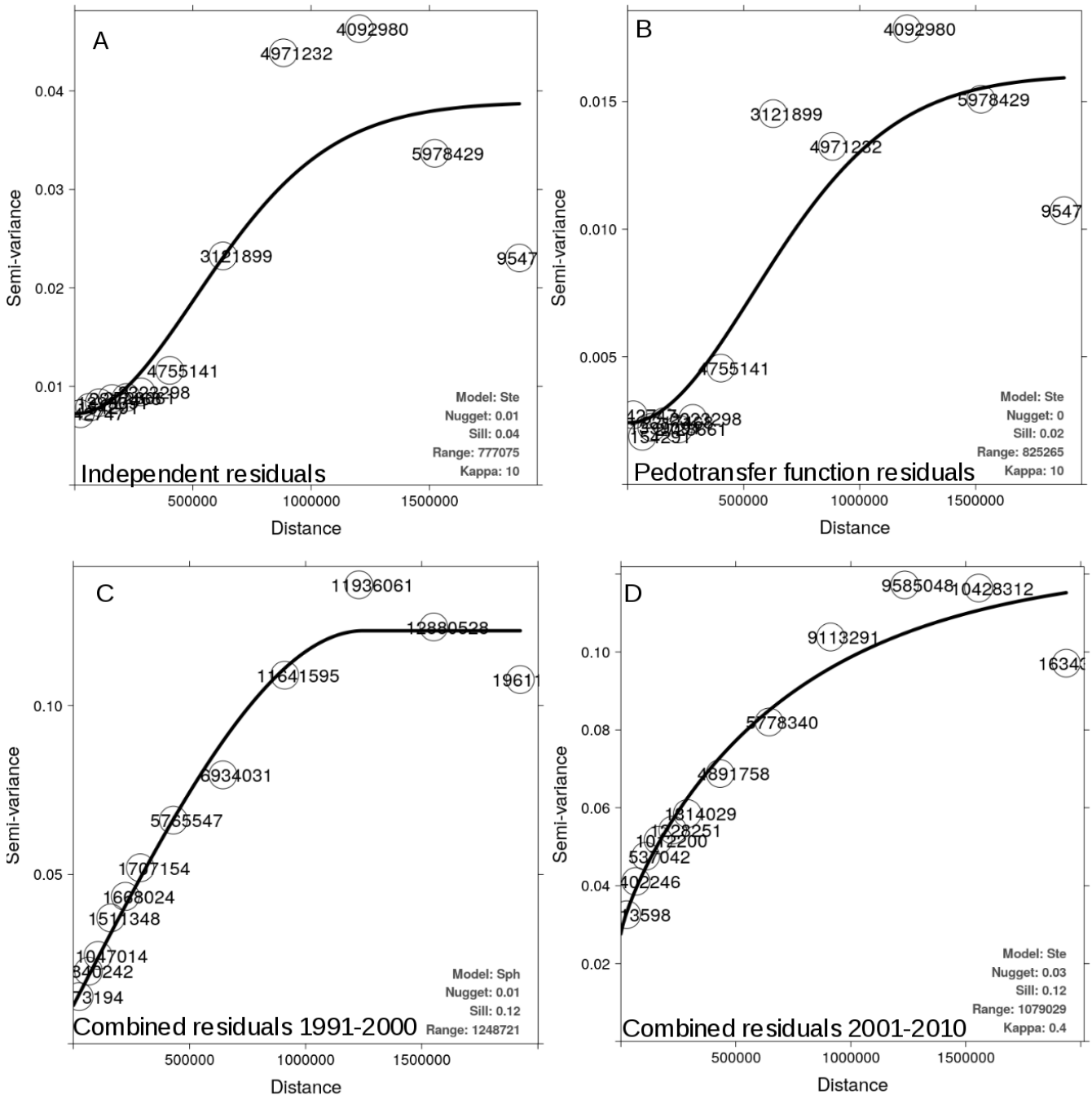


Figure 5 Predicted SOC maps across both countries. The prediction using all data between 1991-2010 (A). The prediction restricting the models to the data between 1991-2001 (B). The prediction restricting the models to the data between 2001-2010 (C).

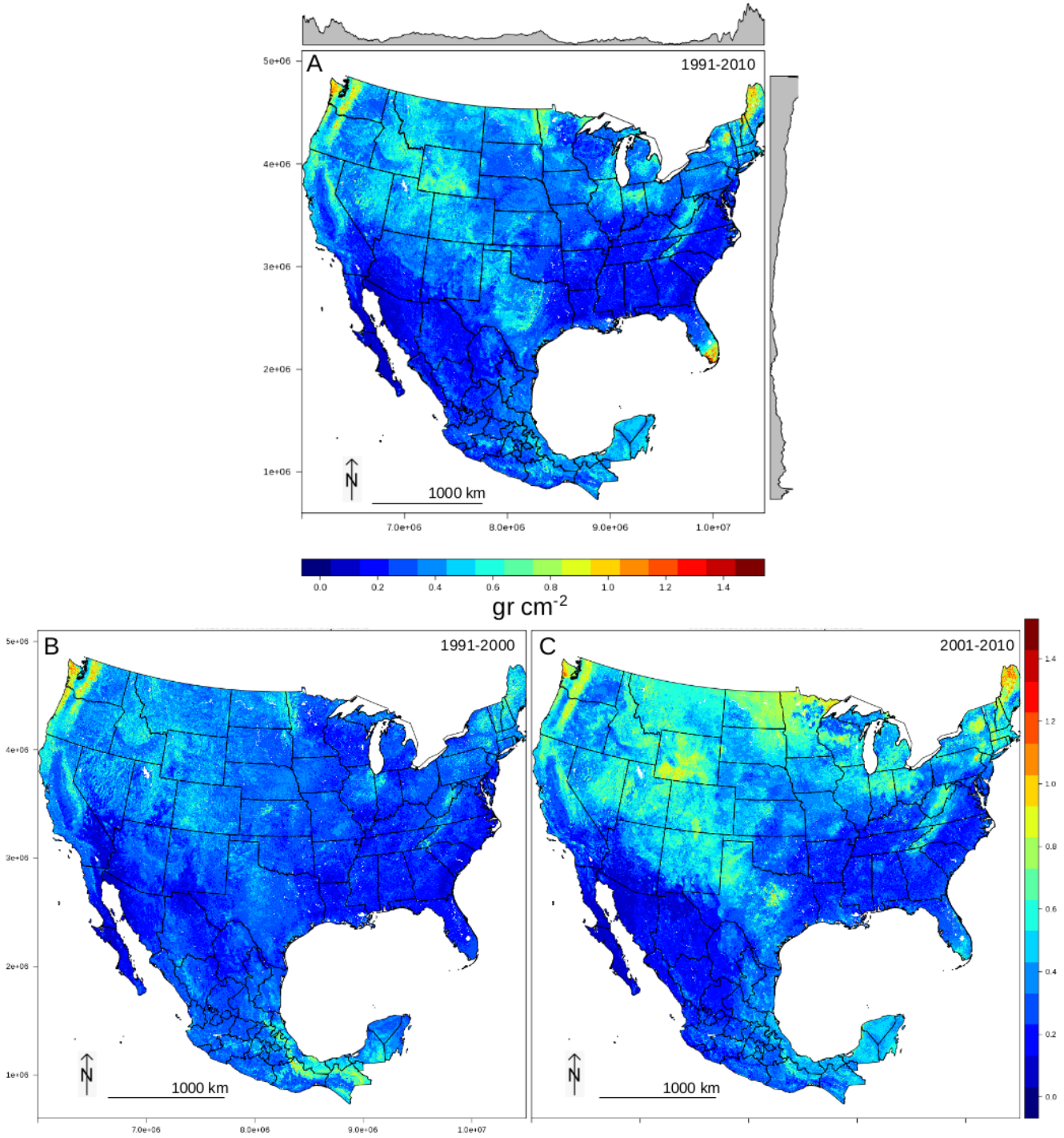


Figure 6 Residual error maps interpolated using Ordinary Kriging to avoid the effect of soil covariates. The variance map of pedotransfer BD residuals (A). The residual error map of our models against independent validation datasets.

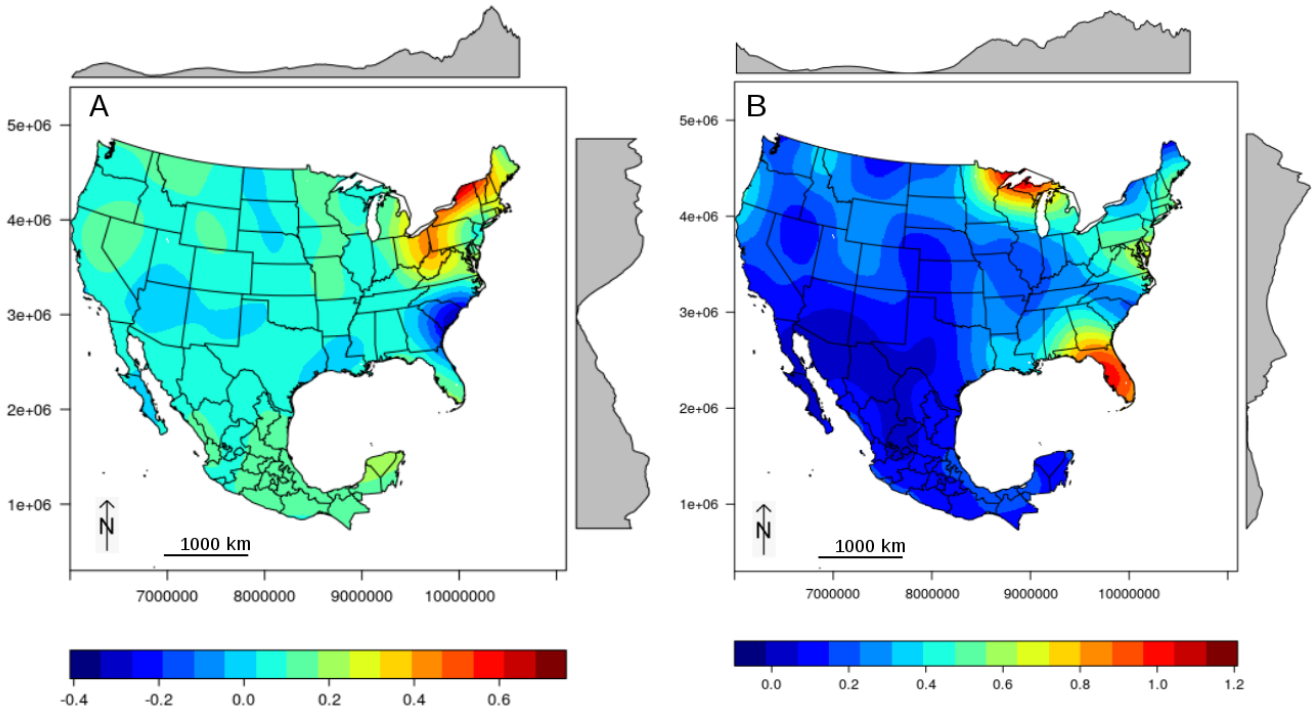


Figure 7 Independent prediction and model variances. The prediction generated using the independent datasets yields 46 Pg of SOC carbon(A). The model variance of our models 1991-2010 (1991-2000 and 2001-2010) (B) and the variance of all models including the independent dataset model (C).

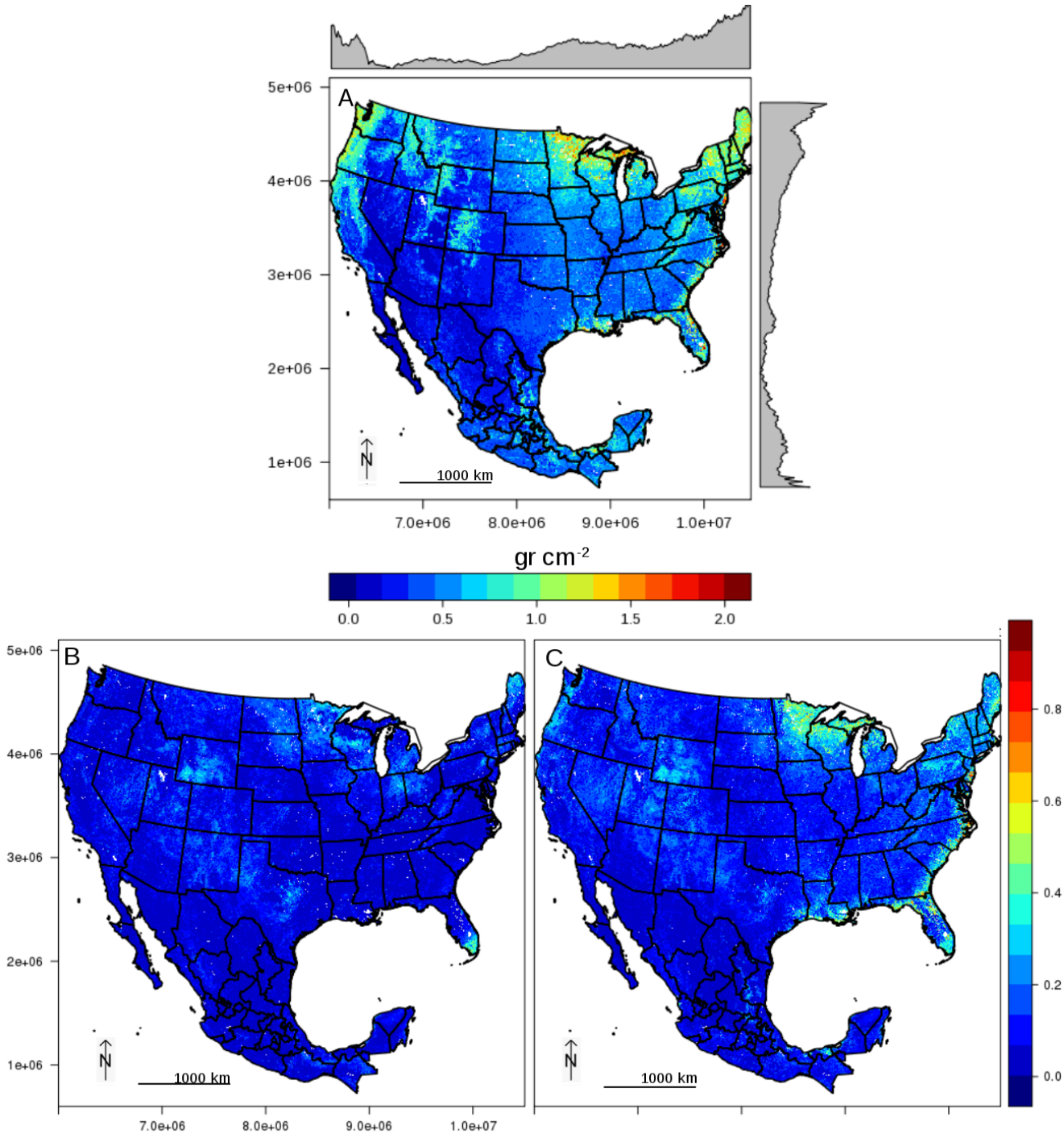
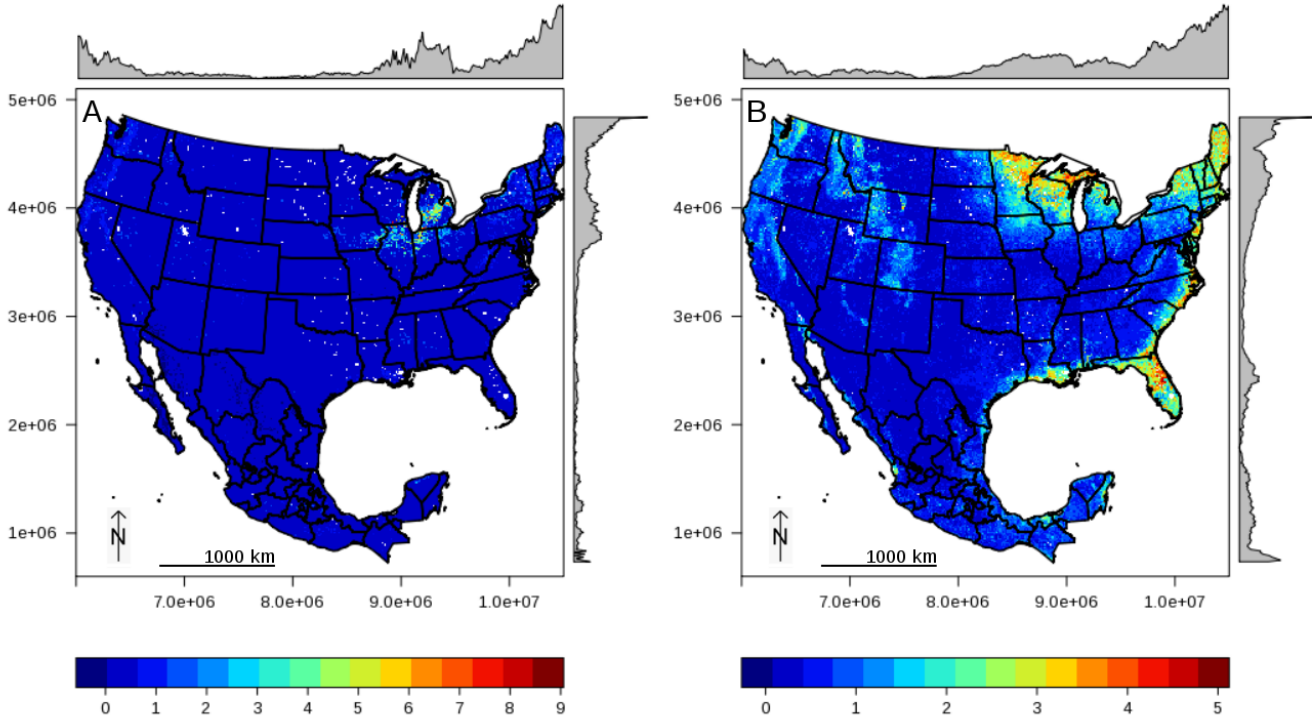
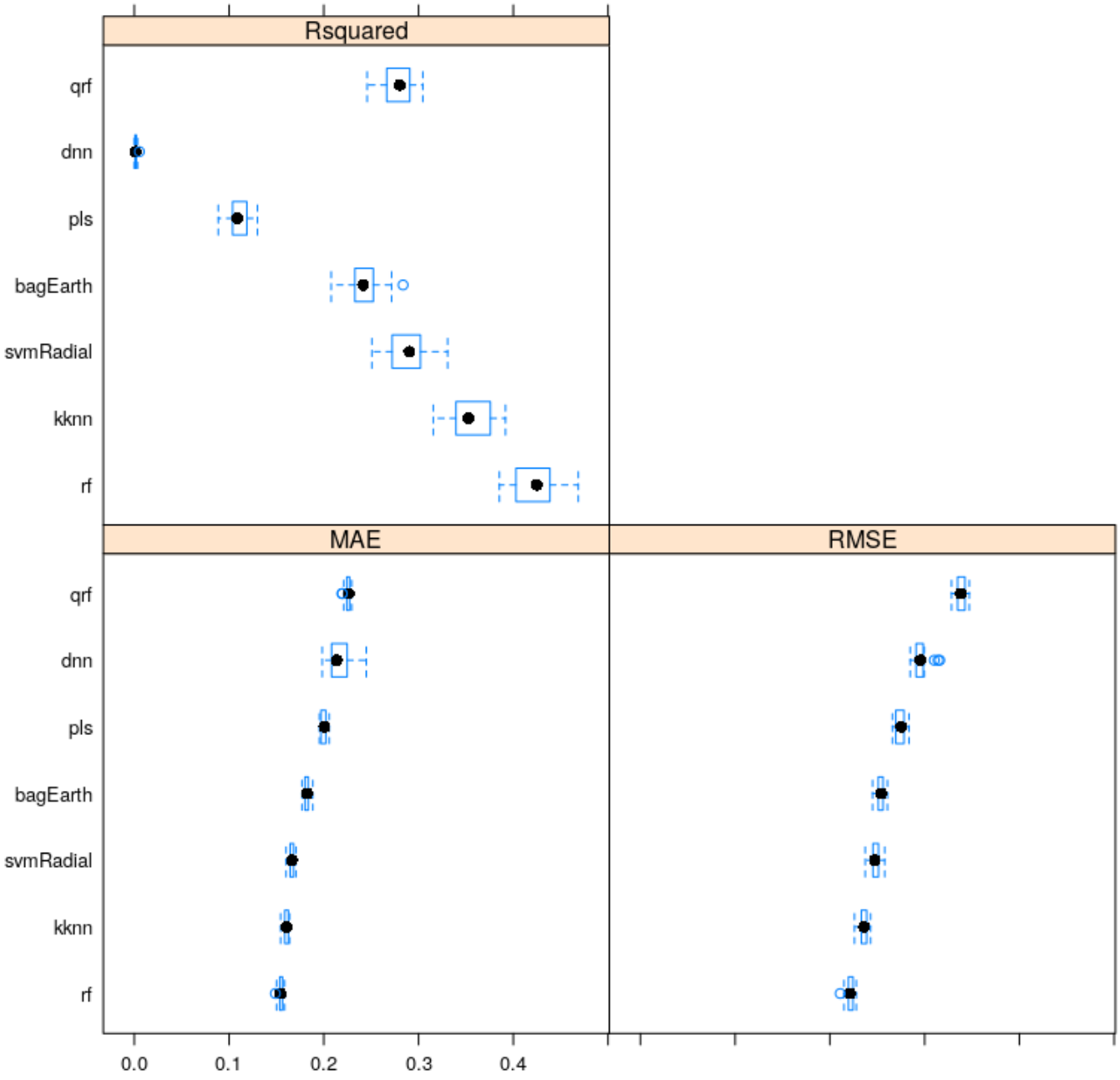


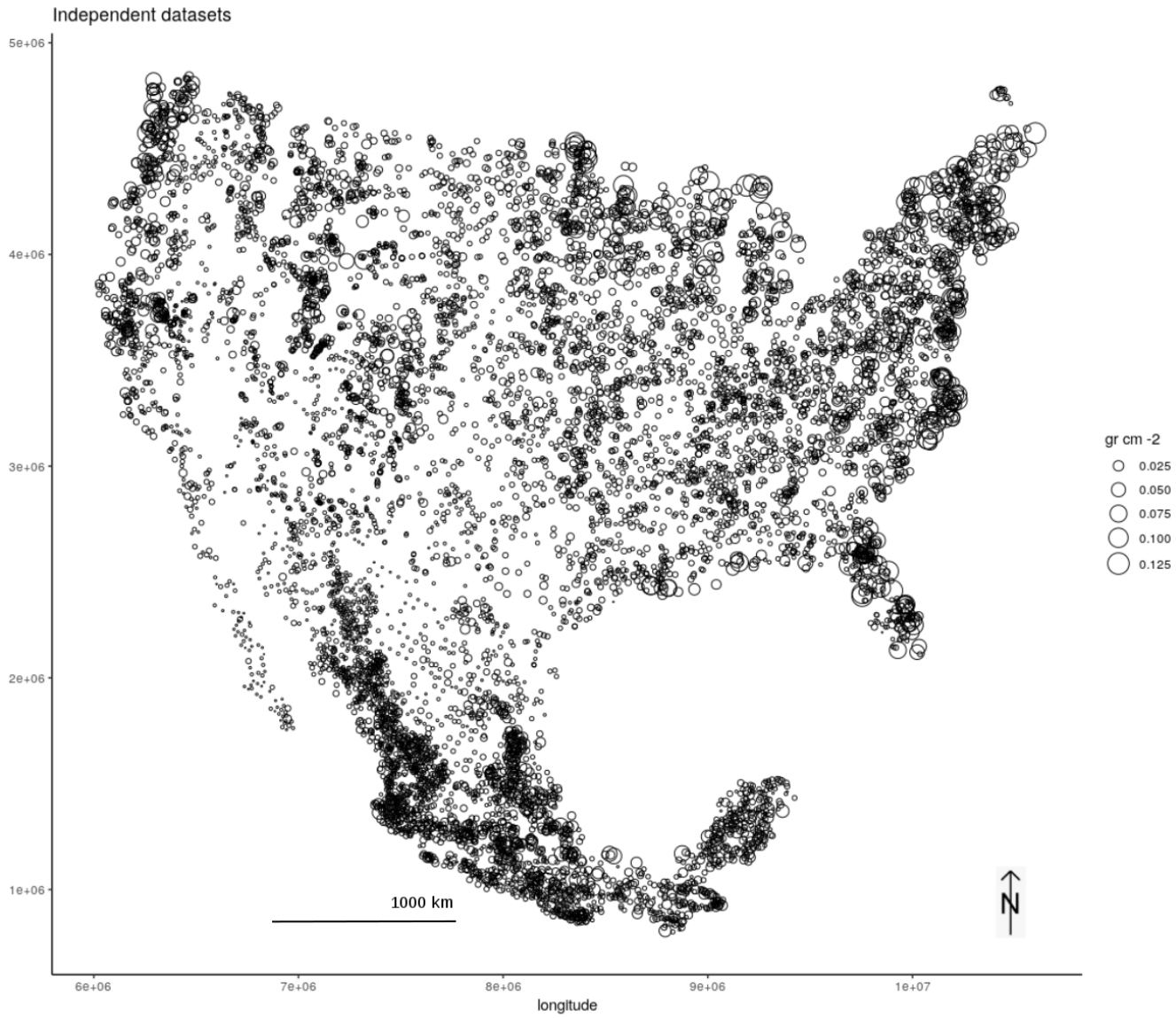
Figure 8 The full conditional distribution of SOC residuals to the best prediction factors. These maps are surrogates of the overall uncertainty given available data and covariates and they are associated to the SOC calculation methods (A) and with the differences against models generated with fully independent datasets (B).



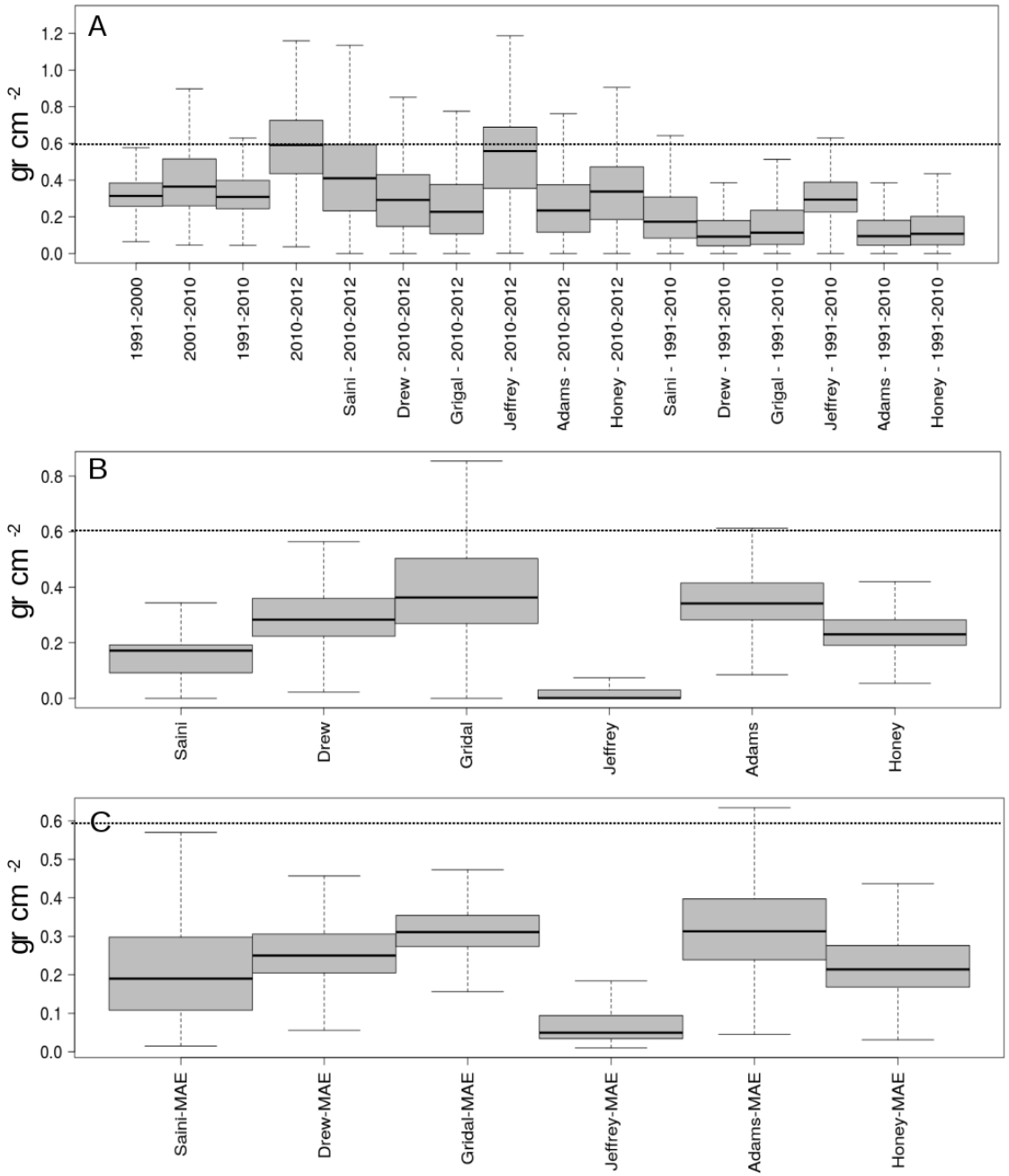
Supplementary Figure S1. Selection of prediction algorithm. Random Forest (rf) generates the lowest error and the highest explained variance. (qrf=quantile regression forest, dnn=deep neural network, pls=partial least squared regression, bagEarth=multivariate adaptive regression splines, svmRadial=radial kernel support vector machines, kknn=kernel weighted nearest neighbors). These results are derived from repeated 5-fold-cross-validation. These methods were implemented using the R package caret. Highest explained variance (Rsquare), lowest mean absolute error (MAE) or root mean squared errors (RMSE) were achieved with rf.



Supplementary Figure S2 Independent datasets (RaCA and the Mexican Forest Service). This combined dataset was collected between 2009-2012.



Supplementary Figure S3 Results from the different peditransfer functions applied to BD for calculating SOC stocks and report the variance of the calculation method. We show the distribution of values in our predictions (1991-2010 first 3 boxplots), the prediction with independent datastes (2010-2012) and the residuals against each of the six pedotransfer functions for BD (A). The estimated BD data from the six epdotransfer functions (B). The mean absolute errors for each SOC estimated value based on the Truncate Taylor series(C). The horizontal line is an arbitrary reference for comparing the values in the three panels.



Supplementary Figure S4 Linear ensemble predicting nearly 50% of SOC variability(Rsquared) using our models as explanatory variables for the independent datasets. These results were cross-validated (5 repeats five folds). The mean absolute error (MAE) and the root mean squared error (RMSE) are betweenn 0.1 and 0.2 gr cm $^{-2}$, the lowest prediction error obtained for SOC data. This ensemble was performed using a random forest (rf) models and a kernel based model (kknn), since were the best approaches predicting SOC data from supplementary Figure S1.