Downscaling Groundwater Storage Data in China to a 1-km Resolution Using Machine Learning Methods

Jianxin Zhang [1,2,3], Kai Liu [1,2,3,*] *and* Ming Wang [1,2,3]

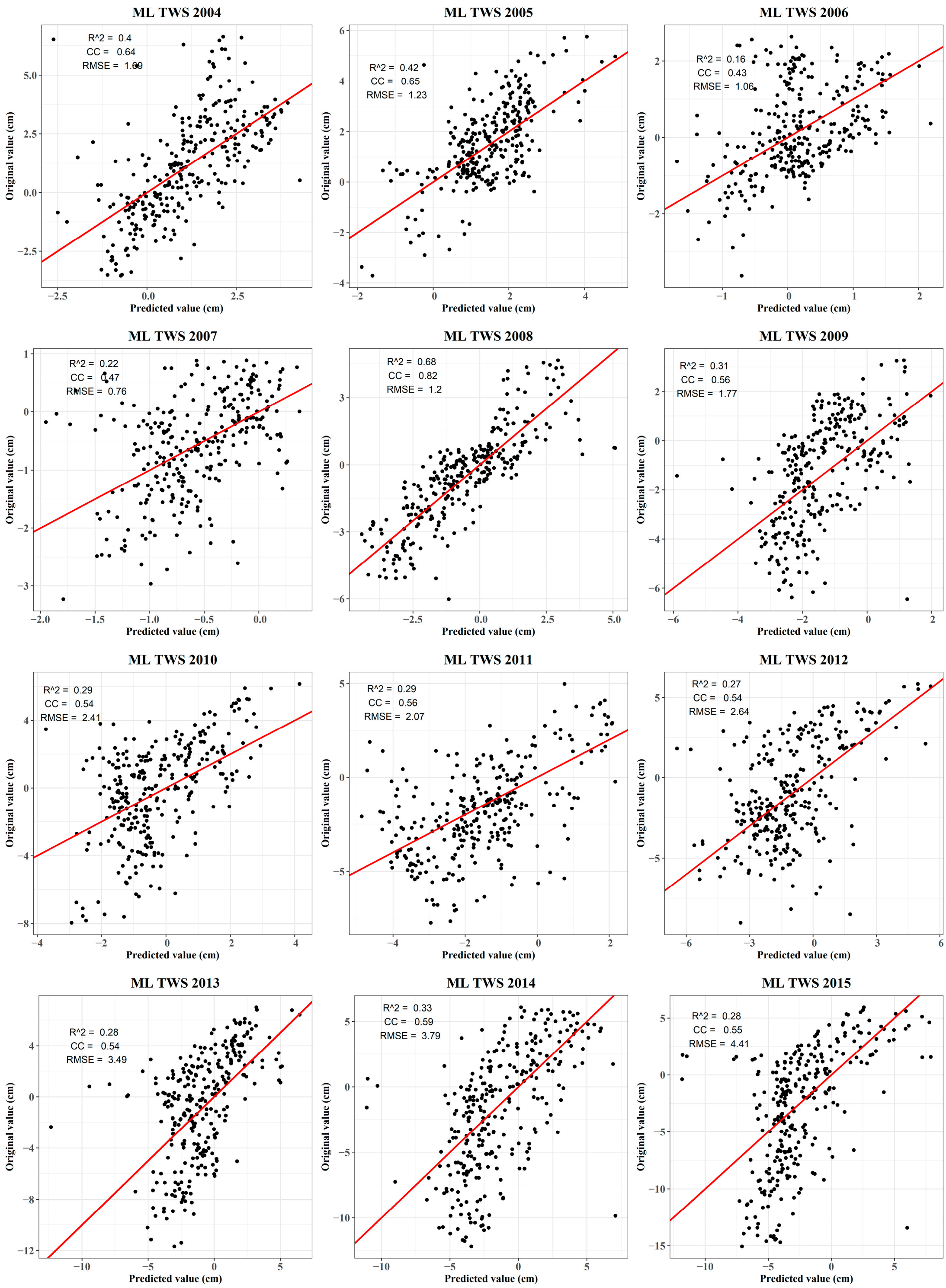
[1] *State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, 100875 Beijing, China*

[2] *Key Laboratory of Environmental Change and Natural Disaster, Beijing Normal University, Beijing 100875, China*

[3] *Academy of Disaster Reduction and Emergency Managemen, Beijing 100875, China*

* Address correspondence to Kai Liu, State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, #19 Xinjiekou Wai Ave., Beijing 100875, China; liukai@bnu.edu.cn

Before performing machine learning methods, we have tried multiple linear (ML) regression methods. Figures S1 and S2 give the scatter plots and the accuracy metrics for each year based on numerical variables. We found that the R^2 ranged from 0.16−0.68 for TWS and 0.09−0.46 for GLDAS. The correlation coefficient (CC) ranged from 0.43−0.82 for TWS and 0.30−0.68 for GLDAS, whilst the RMSE ranged from 0.76−5.41 for TWS and 1.00−2.98 for GLDAS. The accuracy of the ML models in each year is low except 2008, and it can not be used for downscaling.



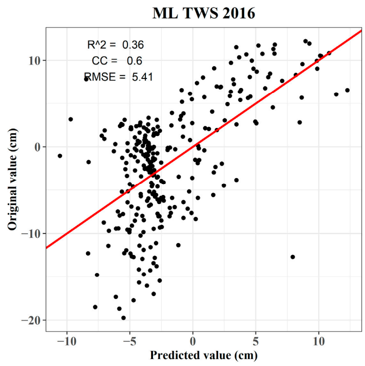
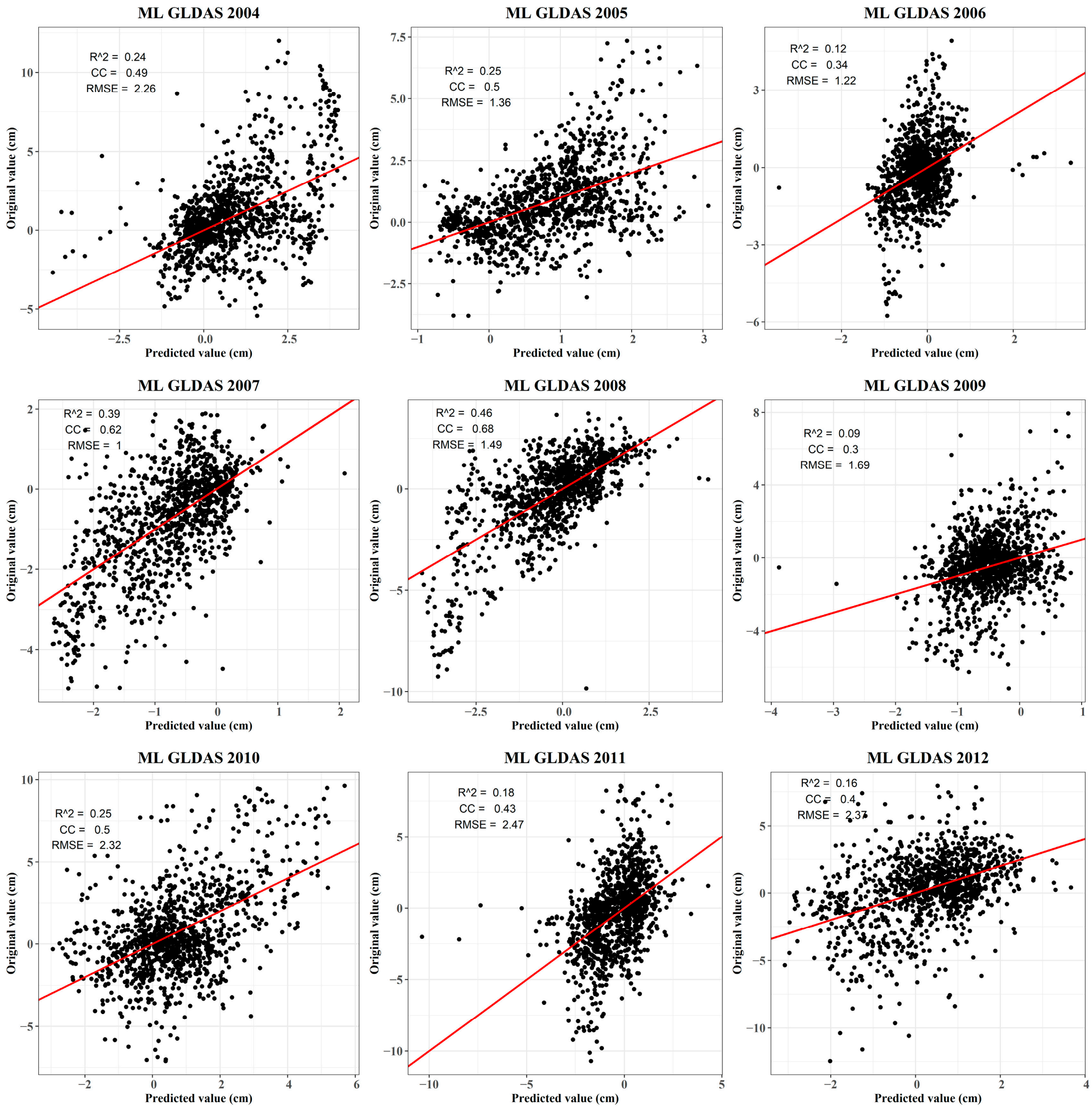


Figure S1. Scatterplots and regression fits of the testing dataset of TWS using multiple linear regression and the original dataset from 2004-2016.



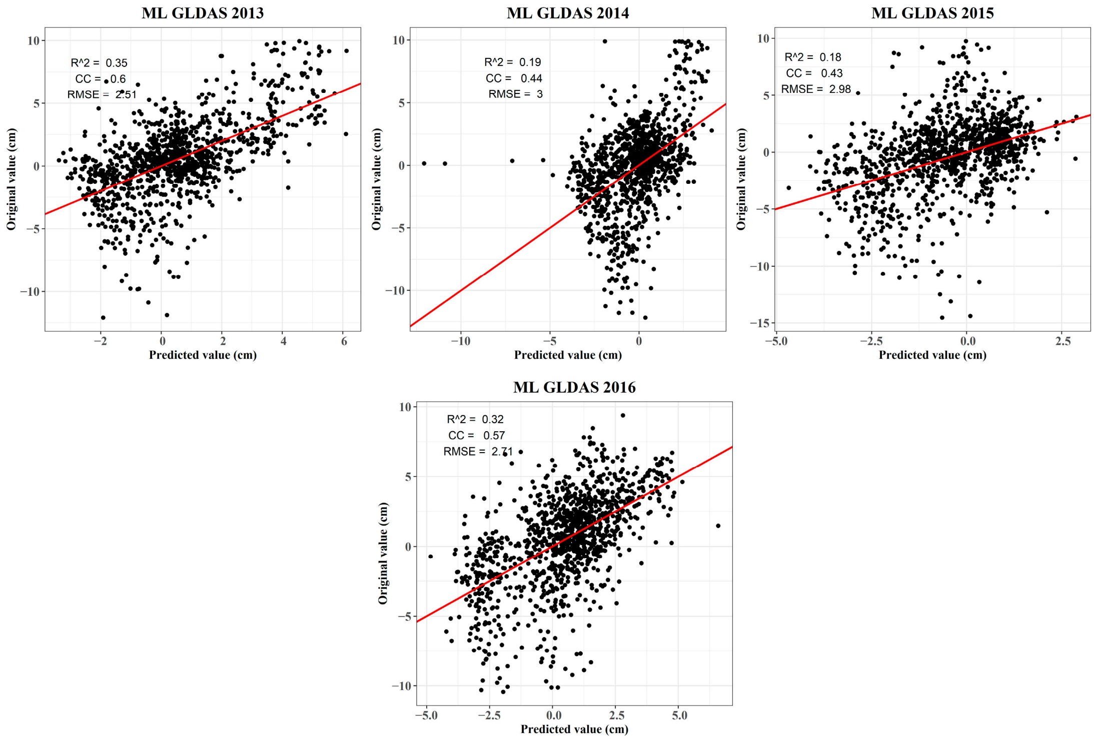


Figure S2. Scatterplots and regression fits of the testing dataset of GLDAS using multiple linear regression and the original dataset from 2004-2016.

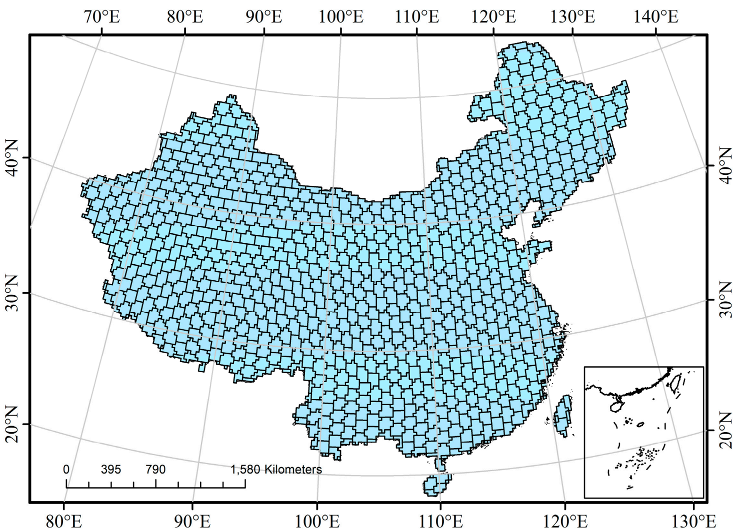
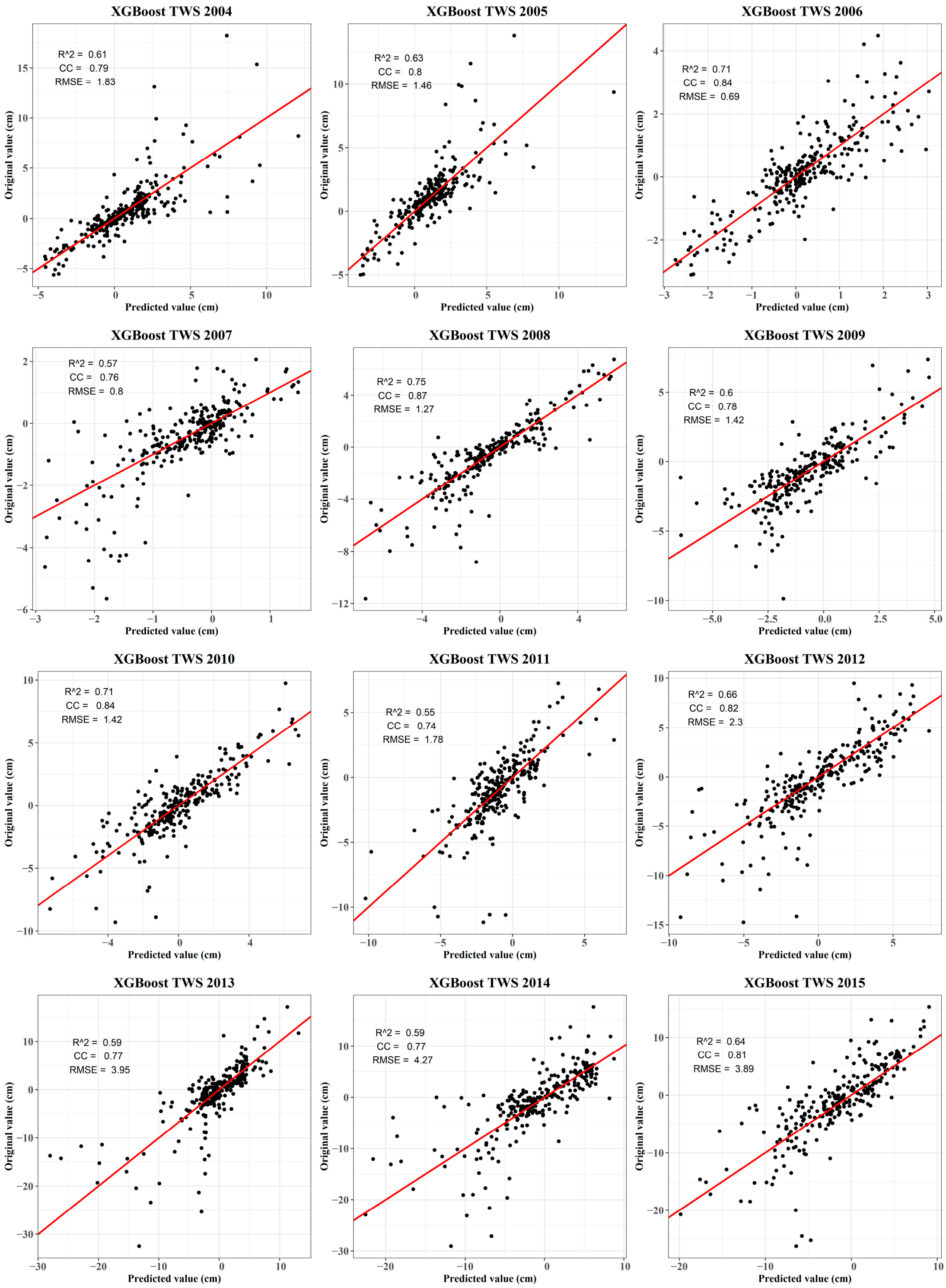


Figure S3 The spatial distribution of 906 clusters using mean shift clustering analysis based on monthly CSR-MS TWS.


XGBoost TWS 2004
R^2 = 0.61
CC = 0.79
RMSE = 1.83
XGBoost TWS 2005
R^2 = 0.63
CC = 0.8
RMSE = 1.46
XGBoost TWS 2006
R^2 = 0.71
CC = 0.84
RMSE = 0.69
XGBoost TWS 2007
R^2 = 0.57
CC = 0.76
RMSE = 0.8
XGBoost TWS 2008
R^2 = 0.75
CC = 0.87
RMSE = 1.27
XGBoost TWS 2009
R^2 = 0.6
CC = 0.78
RMSE = 1.42
XGBoost TWS 2010
R^2 = 0.71
CC = 0.84
RMSE = 1.42
XGBoost TWS 2011
R^2 = 0.55
CC = 0.74
RMSE = 1.78
XGBoost TWS 2012
R^2 = 0.66
CC = 0.82
RMSE = 2.3
XGBoost TWS 2013
R^2 = 0.59
CC = 0.77
RMSE = 3.95
XGBoost TWS 2014
R^2 = 0.59
CC = 0.77
RMSE = 4.27
XGBoost TWS 2015
R^2 = 0.64
CC = 0.81
RMSE = 3.89
Original value (cm)
Predicted value (cm)

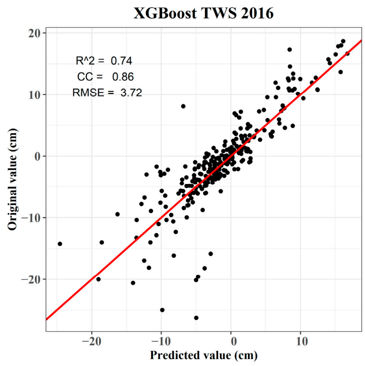
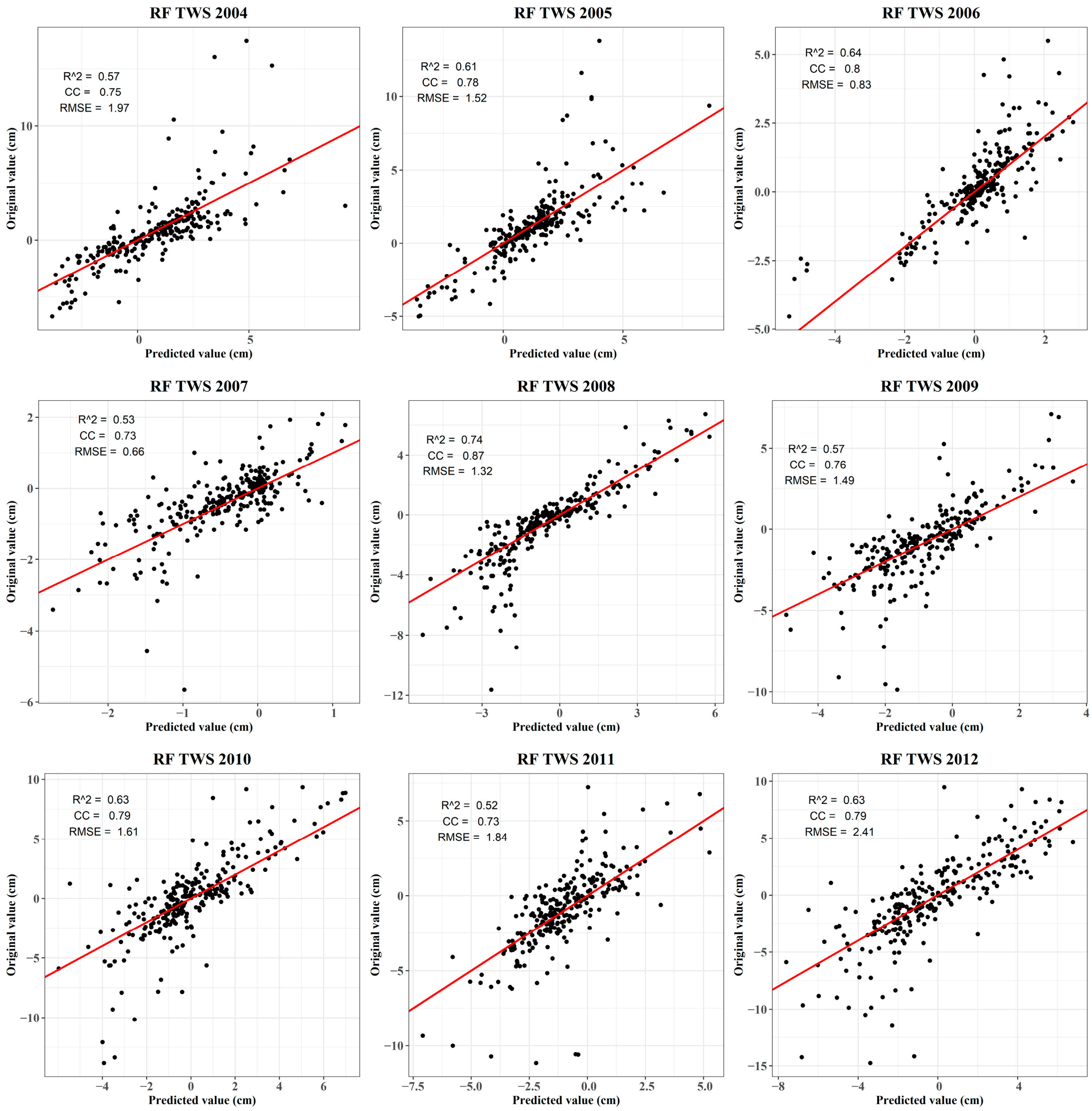


Figure S4 Scatterplots and regression fits of the testing dataset of TWS using XGBoost and the original dataset for 2004–2016 based on CSR-MS 906 clusters analysis.



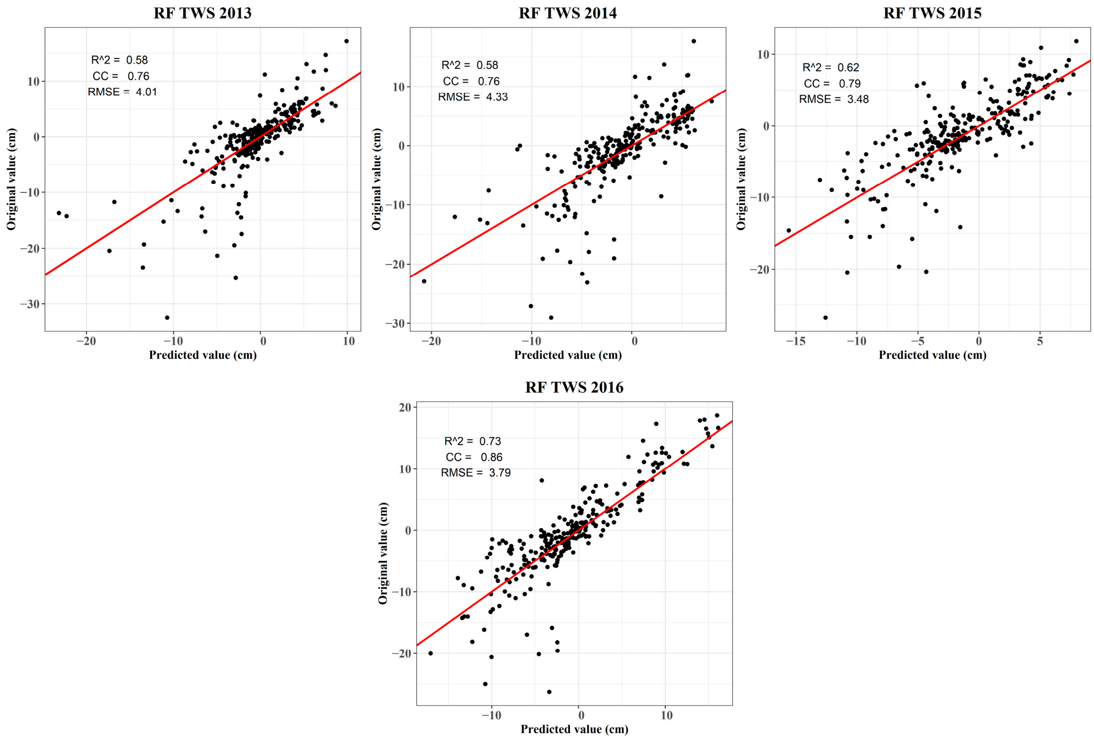
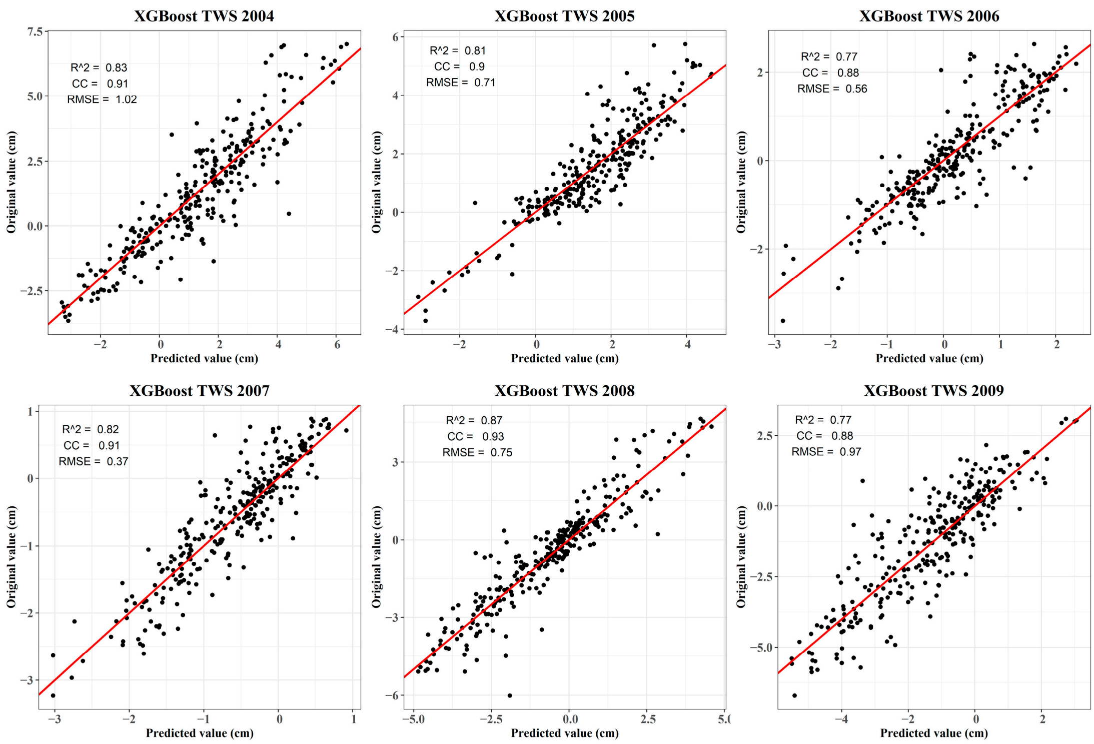


Figure S5 Scatterplots and regression fits of the testing dataset of TWS using RF and the original dataset for 2004–2016 based on CSR-MS 906 clusters analysis.

Figures S6–S13 give scatterplots and variable importance measures for TWS and GLDAS based on RF and XGBoost methods using CSR TELLUS GRACE data for 2004-2016.



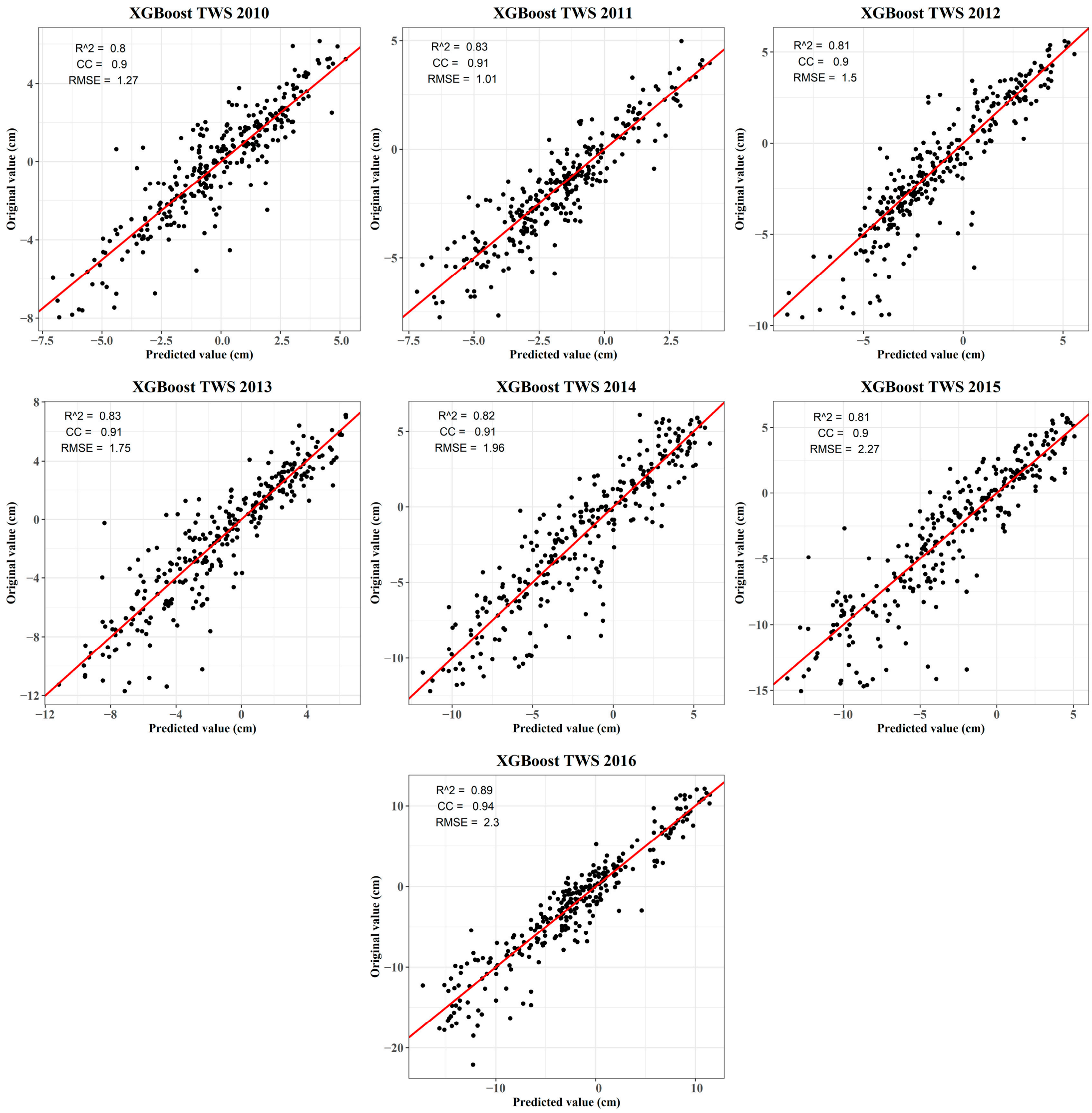
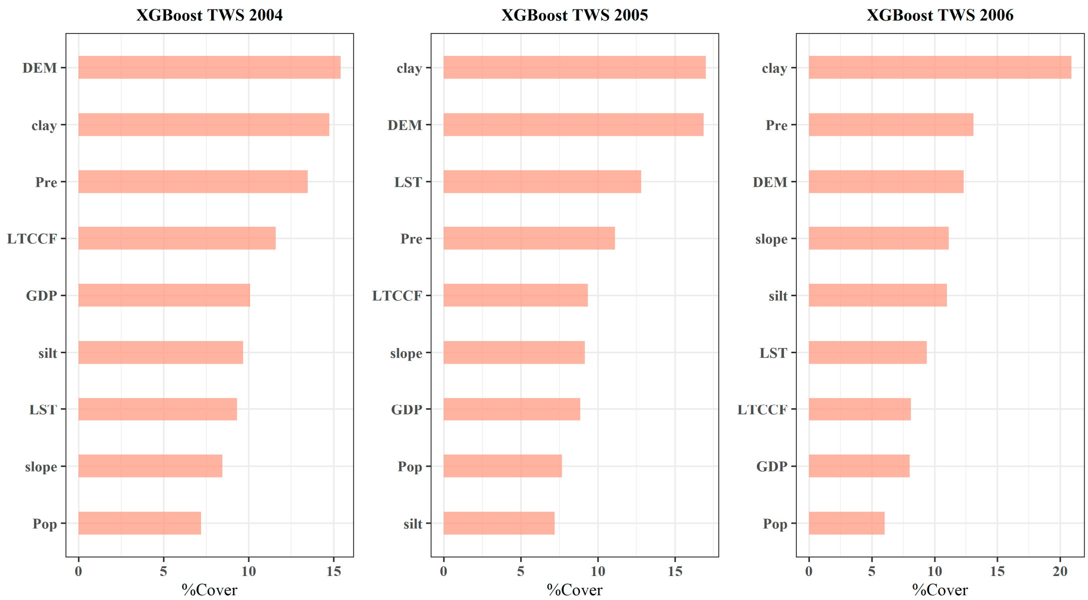
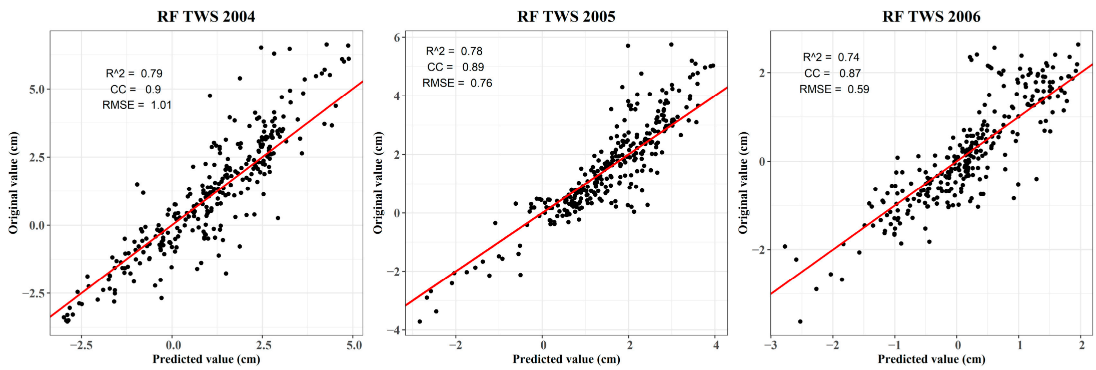


Figure S6 Scatterplots and regression fits of the testing dataset of TWS using XGBoost method.




XGBoost TWS 2007
Pre
silt
DEM
clay
LST
GDP
LTCCF
slope
Pop
0
5
10
%Cover
XGBoost TWS 2008
Pre
LST
DEM
LTCCF
slope
clay
Pop
silt
GDP
0
5
10
15
%Cover
XGBoost TWS 2009
LST
DEM
silt
Pre
slope
clay
GDP
Pop
LTCCF
0
5
10
15
%Cover
XGBoost TWS 2010
silt
clay
DEM
LST
Pre
Pop
LTCCF
slope
GDP
0
5
10
15
%Cover
XGBoost TWS 2011
DEM
Pre
clay
silt
LST
Pop
GDP
slope
LTCCF
0
5
10
15
%Cover
XGBoost TWS 2012
DEM
LST
Pre
GDP
slope
silt
Pop
clay
LTCCF
0
5
10
15
%Cover

Figure S7 Plots of the variable importance measure (Cover metric of the number of observation related to this feature (%Cover)) for the XGBoost-modeled TWS from 2004-2016.



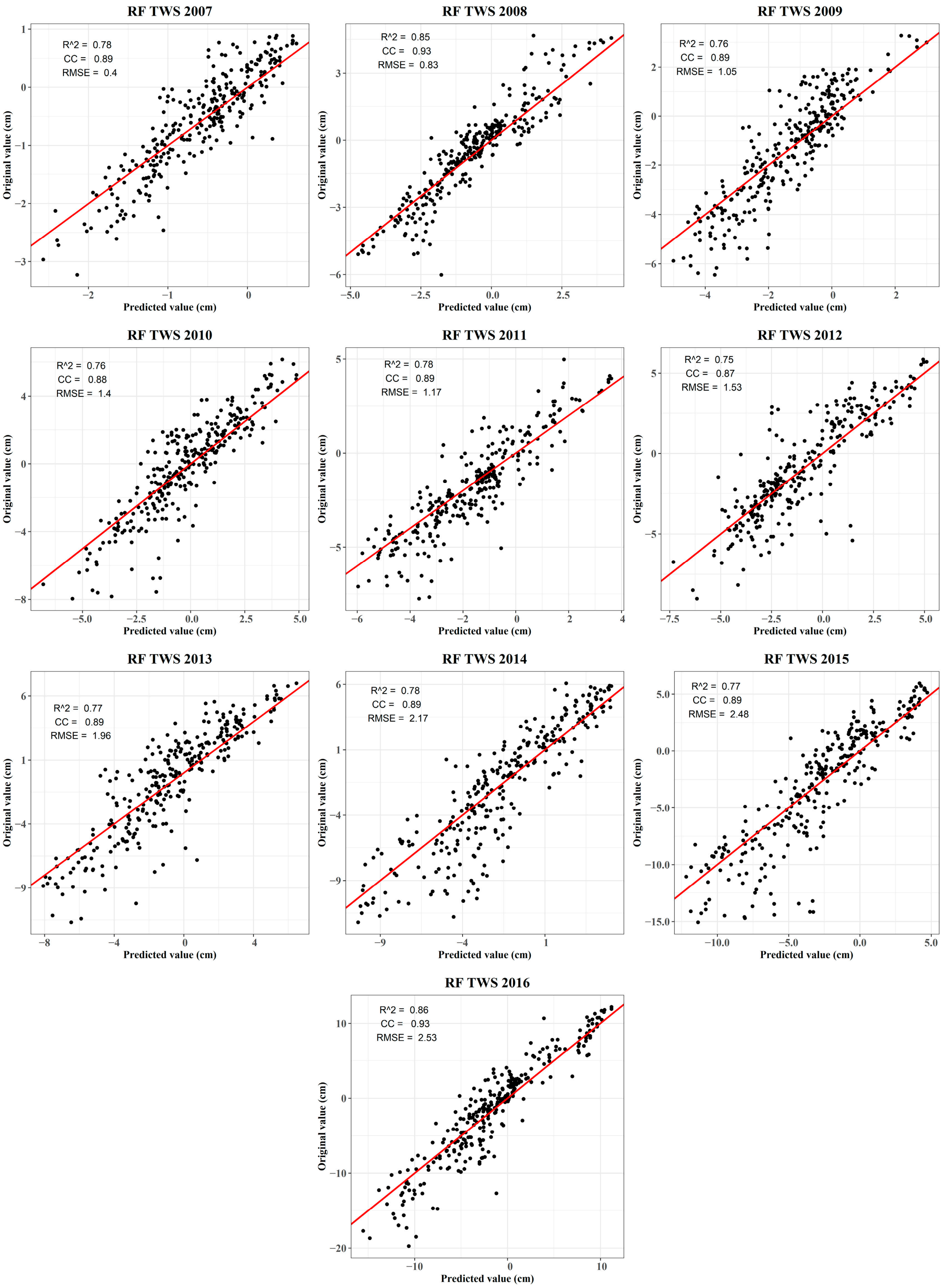
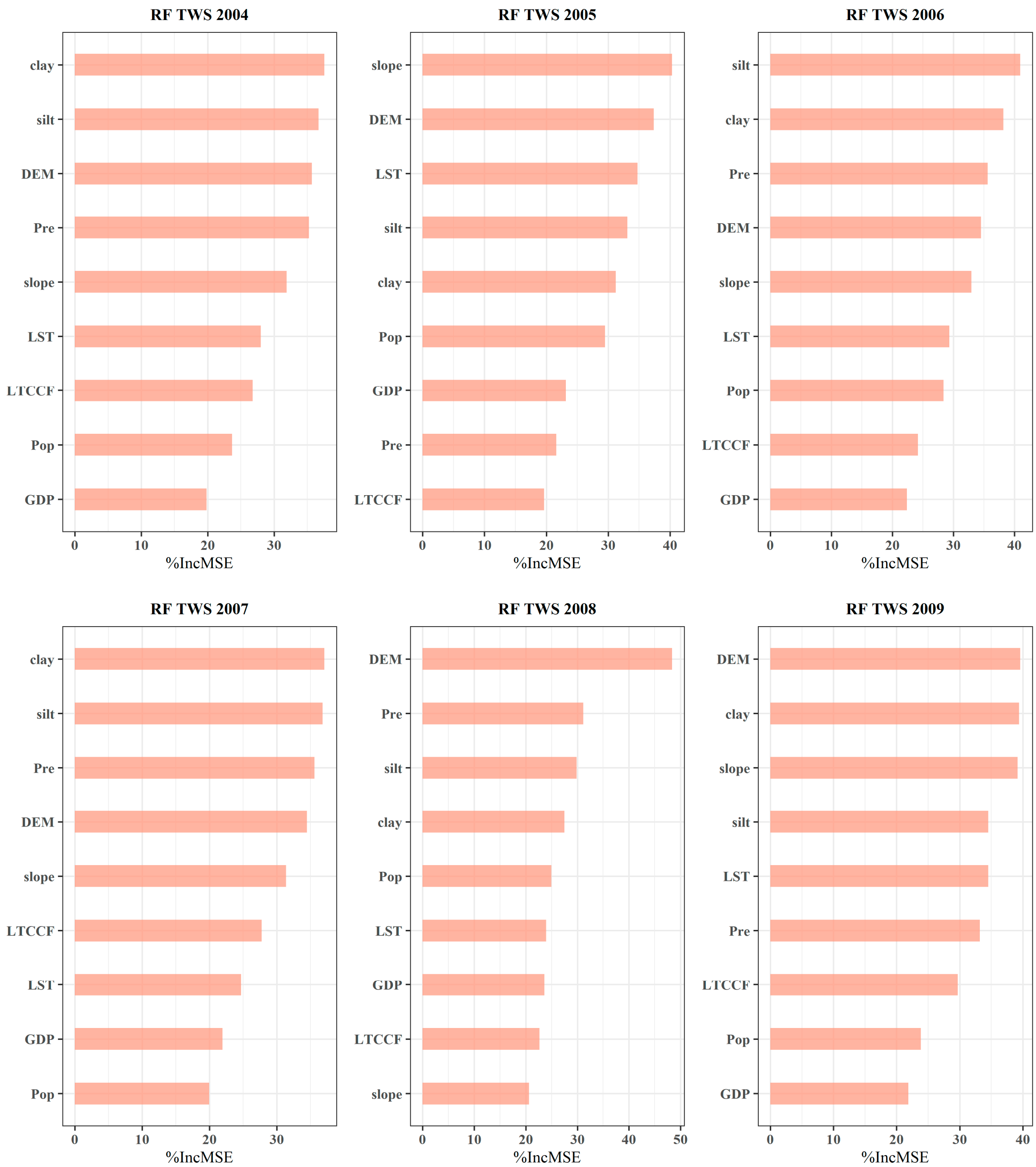
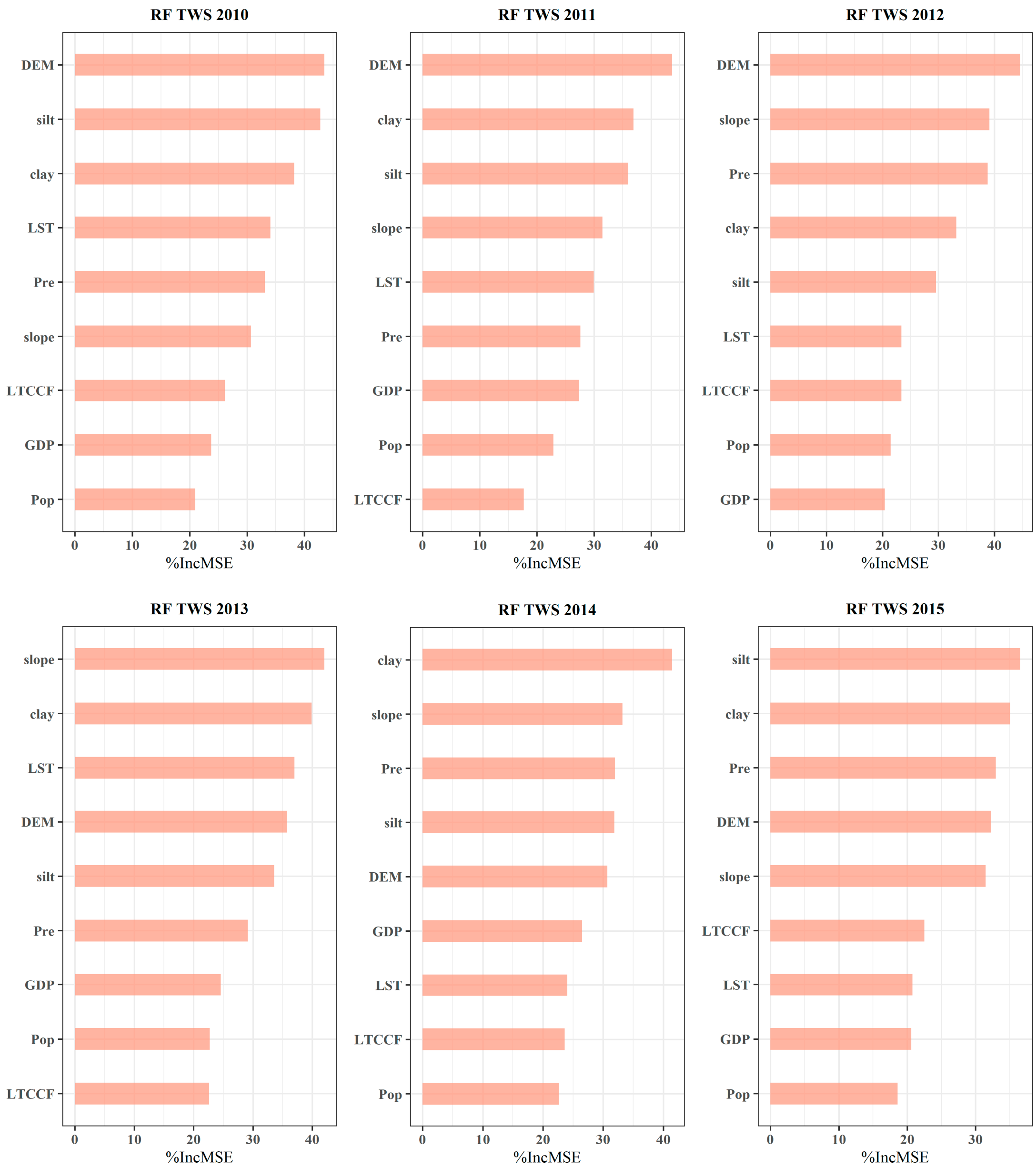


Figure S8 Scatterplots and regression fits of the testing dataset of TWS using RF method.


RF TWS 2004
clay
silt
DEM
Pre
slope
LST
LTCCF
Pop
GDP
0
10
20
30
%IncMSE
RF TWS 2005
slope
DEM
LST
silt
clay
Pop
GDP
Pre
LTCCF
0
10
20
30
40
%IncMSE
RF TWS 2006
silt
clay
Pre
DEM
slope
LST
Pop
LTCCF
GDP
0
10
20
30
40
%IncMSE
RF TWS 2007
clay
silt
Pre
DEM
slope
LTCCF
LST
GDP
Pop
0
10
20
30
%IncMSE
RF TWS 2008
DEM
Pre
silt
clay
Pop
LST
GDP
LTCCF
slope
0
10
20
30
40
50
%IncMSE
RF TWS 2009
DEM
clay
slope
silt
LST
Pre
LTCCF
Pop
GDP
0
10
20
30
40
%IncMSE


RF TWS 2010
DEM
silt
clay
LST
Pre
slope
LTCCF
GDP
Pop
0 10 20 30 40
%IncMSE
RF TWS 2011
DEM
clay
silt
slope
LST
Pre
GDP
Pop
LTCCF
0 10 20 30 40
%IncMSE
RF TWS 2012
DEM
slope
Pre
clay
silt
LST
LTCCF
Pop
GDP
0 10 20 30 40
%IncMSE
RF TWS 2013
slope
clay
LST
DEM
silt
Pre
GDP
Pop
LTCCF
0 10 20 30 40
%IncMSE
RF TWS 2014
clay
slope
Pre
silt
DEM
GDP
LST
LTCCF
Pop
0 10 20 30 40
%IncMSE
RF TWS 2015
silt
clay
Pre
DEM
slope
LTCCF
LST
GDP
Pop
0 10 20 30
%IncMSE

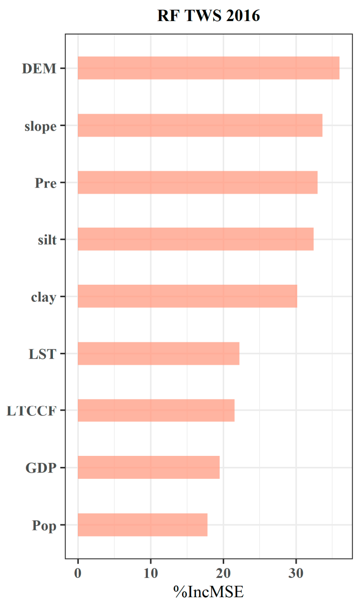
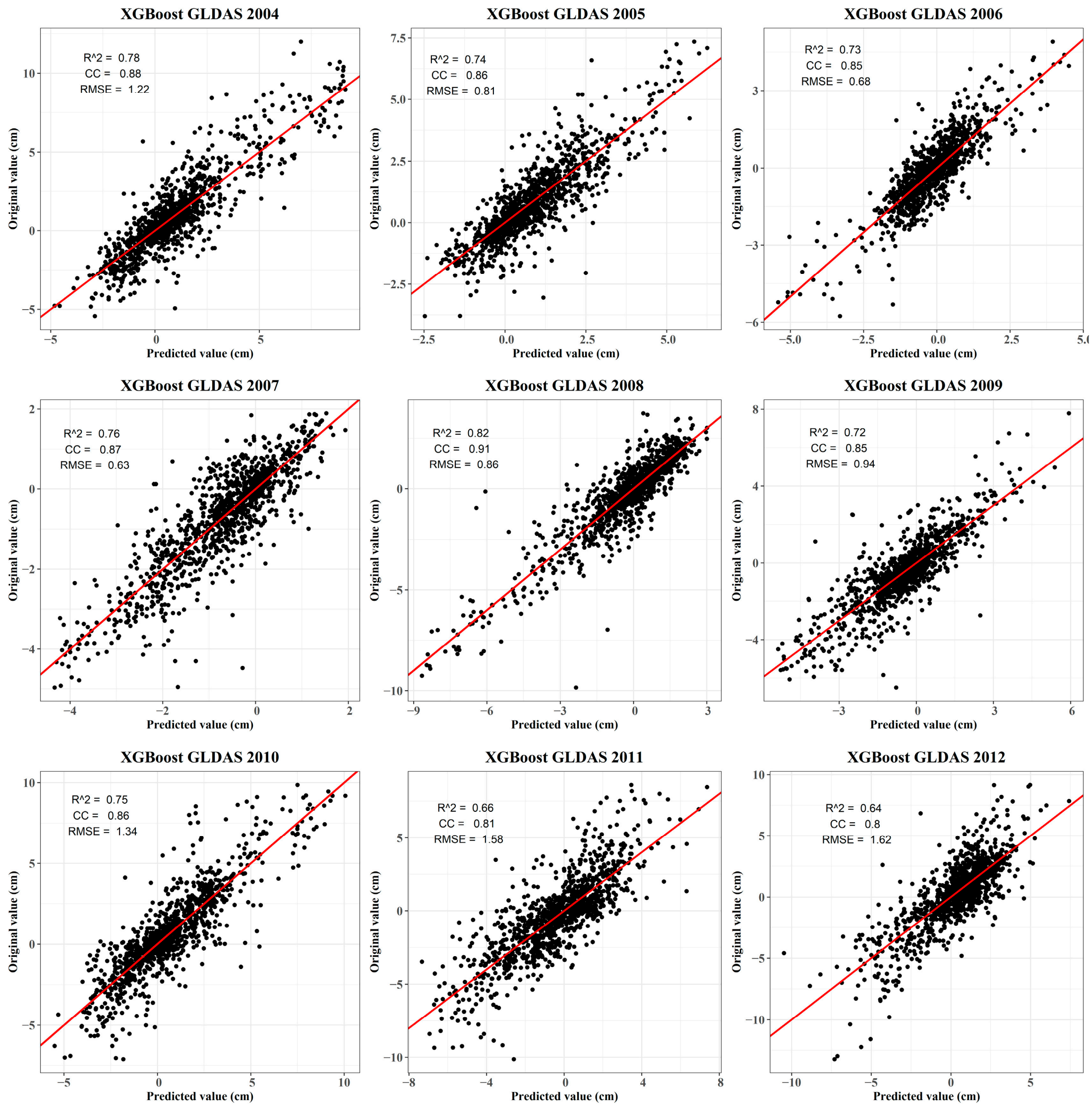


Figure S9 Plots of the variable importance measure (increase in the mean square error (%IncMSE)) for the RF-modeled TWS from 2004-2016.



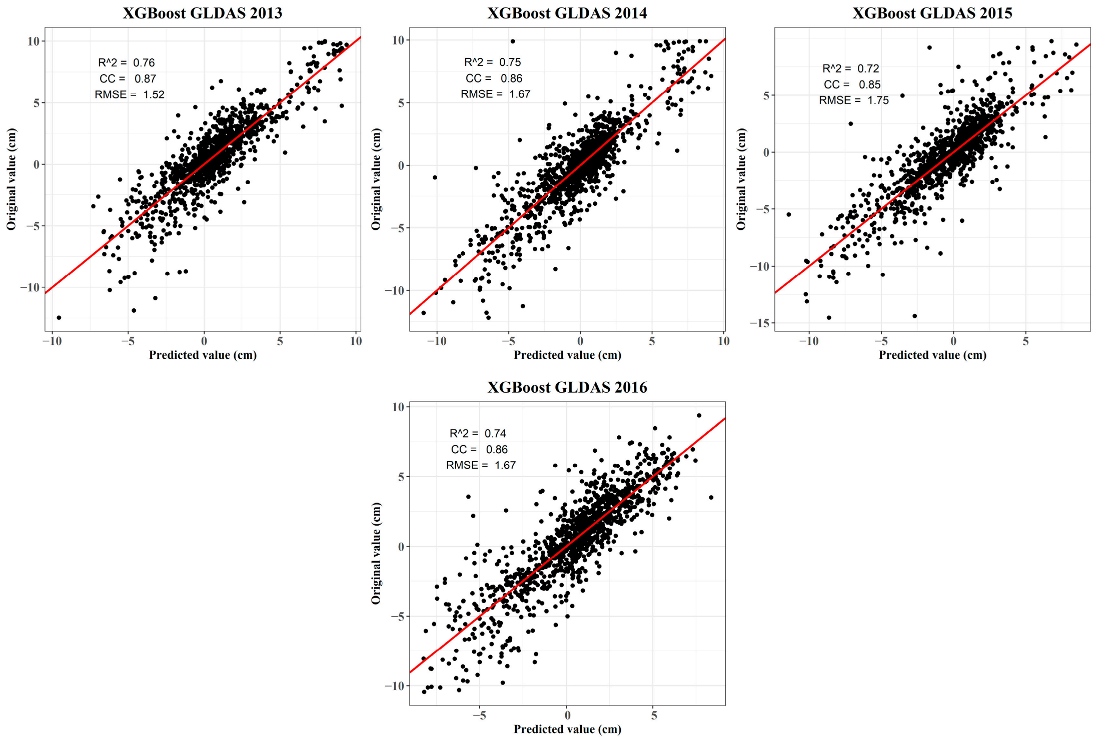
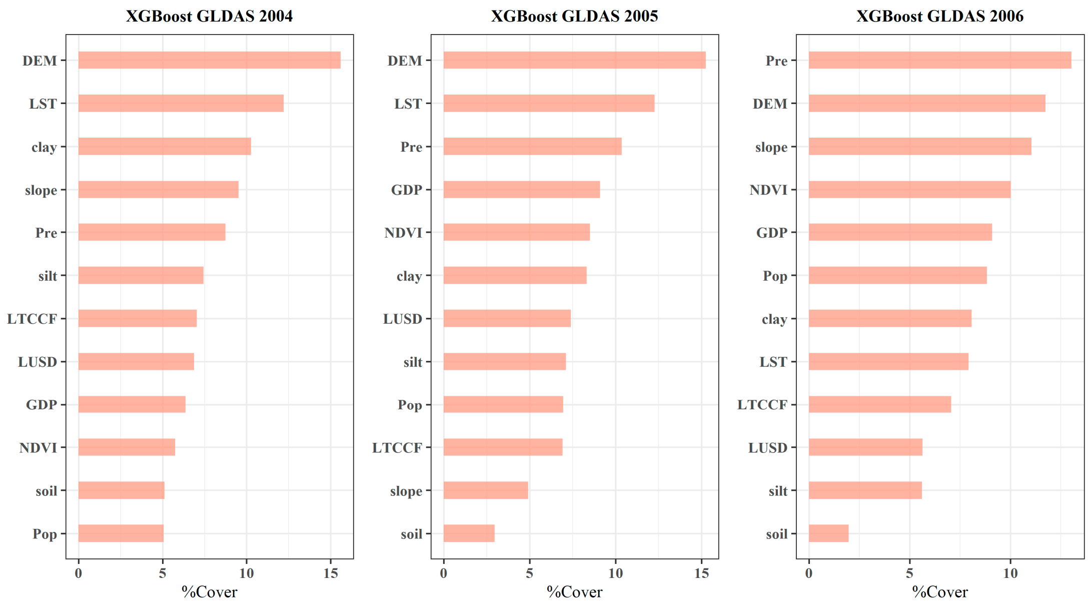
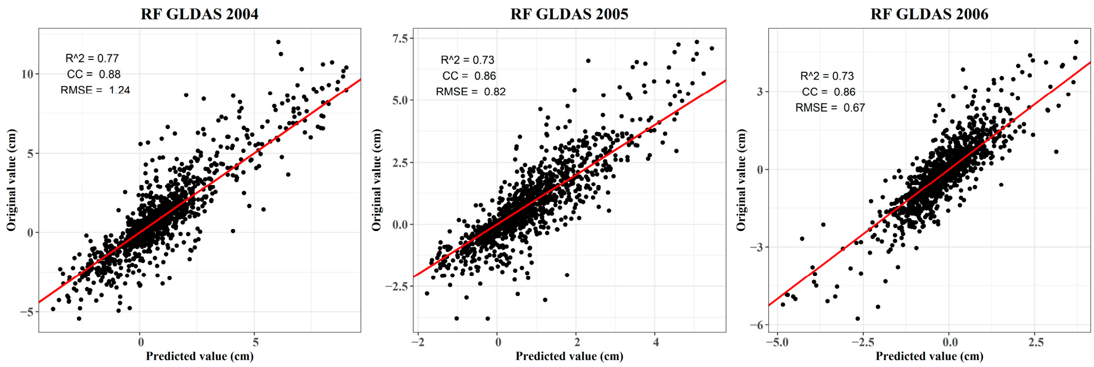


Figure S10 Scatterplots and regression fits of testing dataset of GLDAS using XGBoost method.




XGBoost GLDAS 2007
DEM
Pre
LST
Pop
NDVI
LUSD
silt
slope
clay
GDP
LTCCF
soil
0
5
10
15
%Cover
XGBoost GLDAS 2008
Pre
DEM
LST
slope
clay
NDVI
LTCCF
GDP
Pop
silt
LUSD
soil
0
5
10
15
%Cover
XGBoost GLDAS 2009
DEM
clay
slope
silt
LST
Pre
GDP
LUSD
NDVI
Pop
LTCCF
soil
0
5
10
%Cover
XGBoost GLDAS 2010
clay
DEM
LST
Pre
LUSD
NDVI
GDP
LTCCF
silt
Pop
slope
soil
0
5
10
15
%Cover
XGBoost GLDAS 2011
DEM
Pop
slope
GDP
clay
LST
silt
Pre
NDVI
LUSD
LTCCF
soil
0
5
10
%Cover
XGBoost GLDAS 2012
DEM
LST
NDVI
slope
LTCCF
clay
Pre
silt
GDP
Pop
LUSD
soil
0
5
10
%Cover

Figure S11 Plots of the variable importance measure (Cover metric of the number of observation related to this feature (%Cover)) for the XGBoost-modeled GLDAS from 2004-2016.



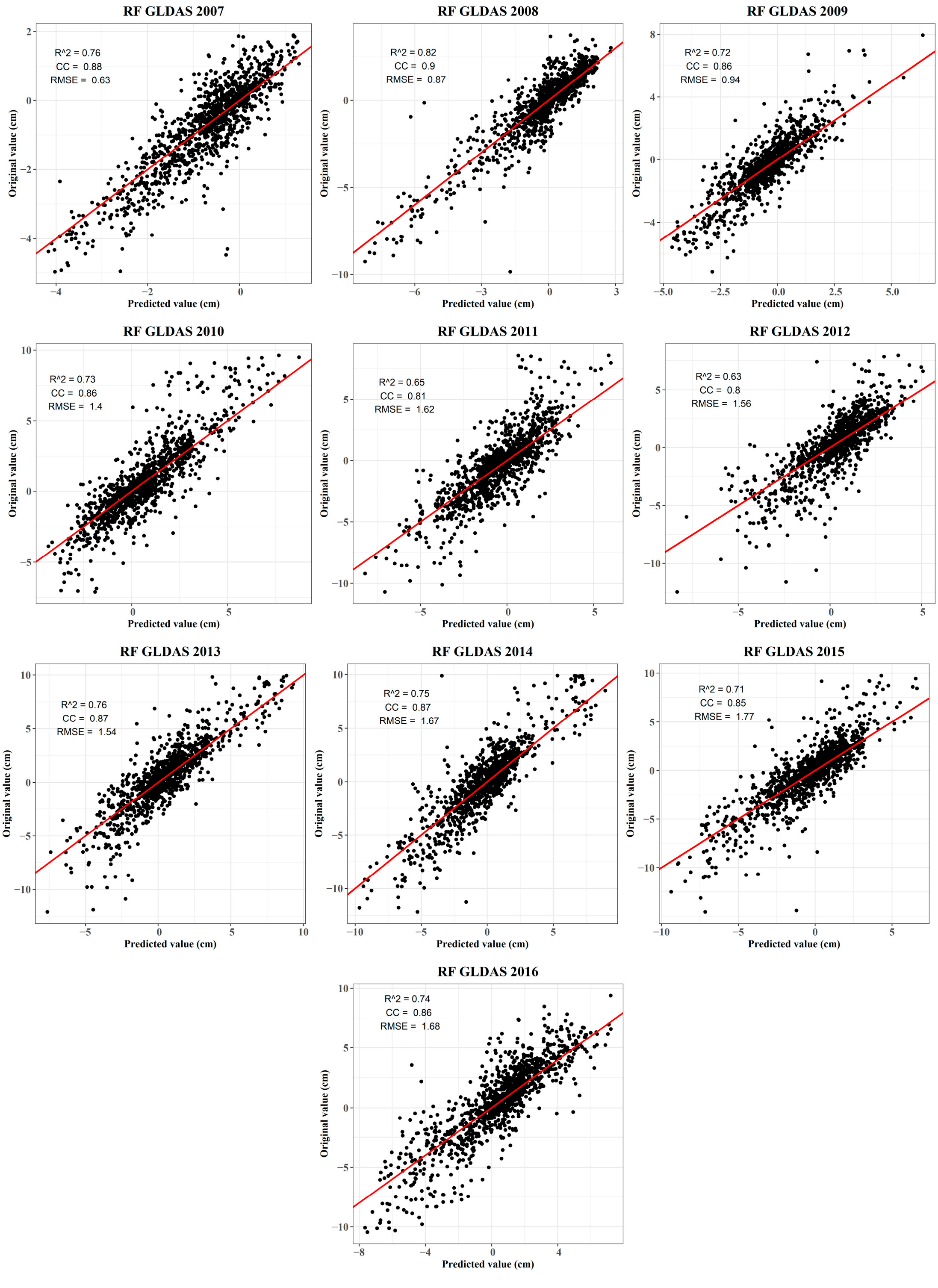
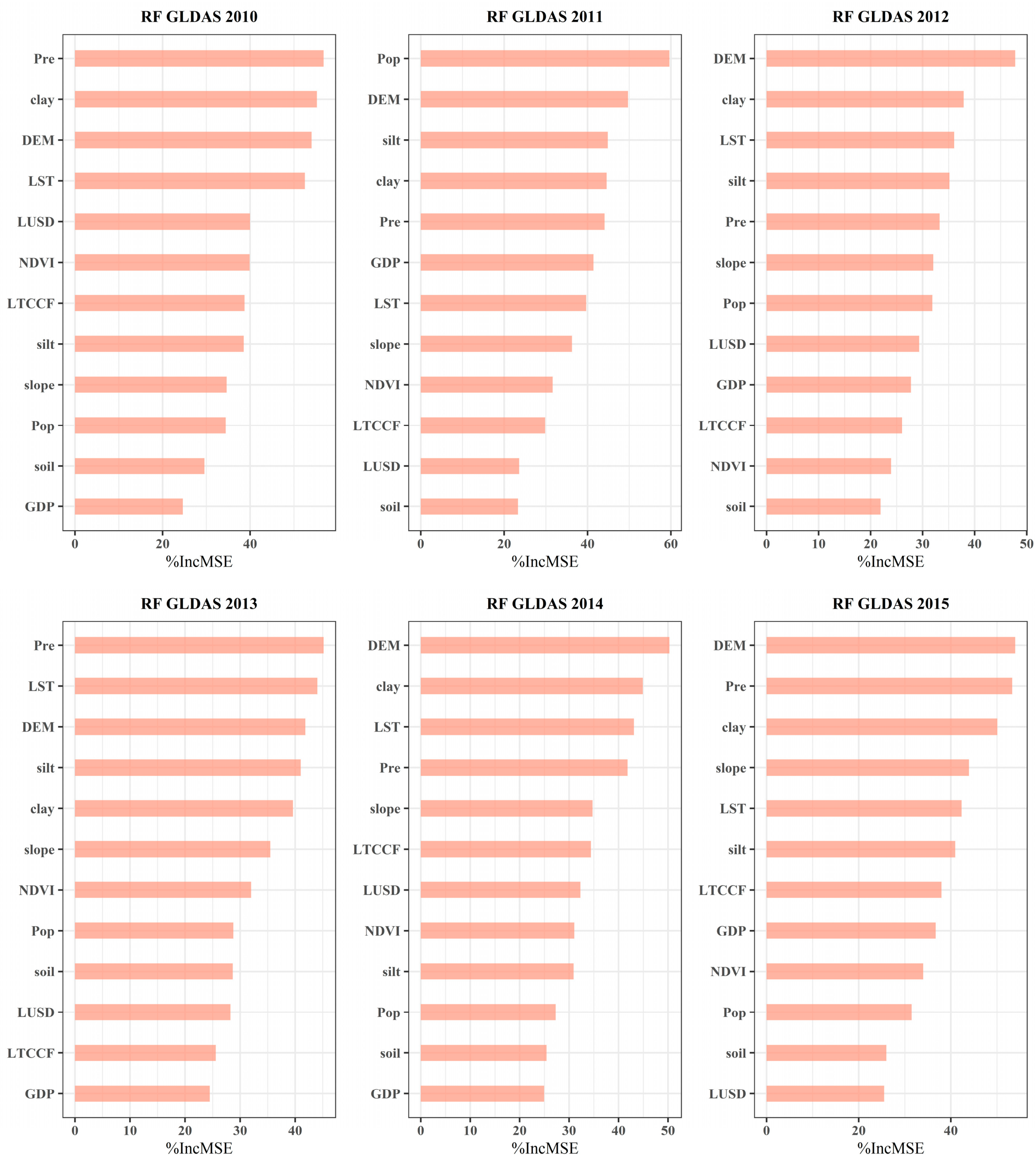


Figure S12 Scatterplots and regression fits of the testing dataset of GLDAS using RF method.


RF GLDAS 2004
LST
Pre
DEM
clay
LTCCF
LUSD
silt
slope
soil
Pop
NDVI
GDP
0
10
20
30
40
50
%IncMSE
RF GLDAS 2005
Pre
DEM
silt
clay
LST
slope
GDP
LTCCF
NDVI
LUSD
Pop
soil
0
20
40
60
%IncMSE
RF GLDAS 2006
DEM
Pre
silt
LST
clay
slope
LTCCF
GDP
LUSD
NDVI
Pop
soil
0
20
40
%IncMSE
RF GLDAS 2007
Pre
clay
DEM
LUSD
silt
slope
GDP
Pop
LST
LTCCF
NDVI
soil
0
20
40
%IncMSE
RF GLDAS 2008
Pre
DEM
LTCCF
clay
GDP
LST
slope
LUSD
silt
NDVI
soil
Pop
0
10
20
30
40
%IncMSE
RF GLDAS 2009
LST
DEM
clay
slope
silt
Pre
LUSD
NDVI
GDP
LTCCF
Pop
soil
0
20
40
60
%IncMSE


RF GLDAS 2010
Pre
clay
DEM
LST
LUSD
NDVI
LTCCF
silt
slope
Pop
soil
GDP
0
20
40
%IncMSE
RF GLDAS 2011
Pop
DEM
silt
clay
Pre
GDP
LST
slope
NDVI
LTCCF
LUSD
soil
0
20
40
60
%IncMSE
RF GLDAS 2012
DEM
clay
LST
silt
Pre
slope
Pop
LUSD
GDP
LTCCF
NDVI
soil
0
10
20
30
40
50
%IncMSE
RF GLDAS 2013
Pre
LST
DEM
silt
clay
slope
NDVI
Pop
soil
LUSD
LTCCF
GDP
0
10
20
30
40
%IncMSE
RF GLDAS 2014
DEM
clay
LST
Pre
slope
LTCCF
LUSD
NDVI
silt
Pop
soil
GDP
0
10
20
30
40
50
%IncMSE
RF GLDAS 2015
DEM
Pre
clay
slope
LST
silt
LTCCF
GDP
NDVI
Pop
soil
LUSD
0
20
40
%IncMSE

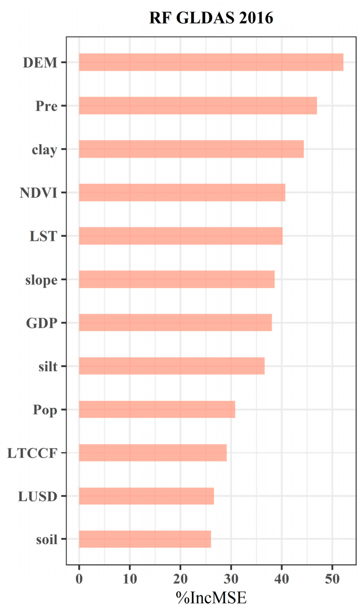


Figure S13 Plots of the variable importance measure (increase in the mean square error (%IncMSE)) for the RF-modeled GLDAS from 2004-2016.