Remote sensing based yield estimation of rice (*Oryza sativa* L.) using gradient boosted regression in India

Ponraj Arumugam, Abel Chemura, Bernhard Schauberger and Christoph Gornott

Table S1: Parameters tested in the hyperparameter tuning to find the best model from GBR decision trees for all the states.

States	Observations	n_estimators	learning_rate	max_depth	min_sample_split	min_sample_leaf
Uttar Pradesh	26775	500–5000 (interval of 500)	0.01–0.001 (interval of 0.001)	3–15 (interval of 2)	2–10 (interval of 2)	1 and 2 –4 (interval of 2)
Punjab	11942					
Chhattisgarh	11035					
West Bengal	10674					
Andhra Pradesh	9847					
Odisha	6678	500–3000 (interval of 500)	0.01–0.001 (interval of 0.001)	3–15 (interval of 2)	2–10 (interval of 2)	1 and 2 –4 (interval of 2)
Madhya Pradesh	6003					
Telangana	5291					
Karnataka	5155					
Maharashtra	4937					
Haryana	3773					
Bihar	2495	100–1900 (interval of 300)	0.1–0.01 (interval of 0.01)	3–13 (interval of 2)	2–10 (interval of 2)	1 and 2 –4 (interval of 2)
Gujarat	2417					
Jharkhand	1264					
Uttarakhand	922					
Rajasthan	651	100–1000 (interval of 100)	0.1–0.01 (interval of 0.01)	3–13 (interval of 2)	2–10 (interval of 2)	1 and 2 –4 (interval of 2)
Jammu & Kashmir	413					
Himachal Pradesh	237					

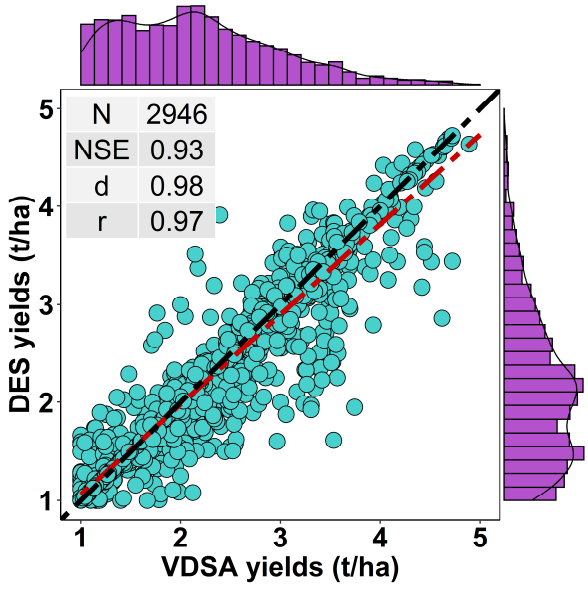

Figure S1. Comparison between two different observed yields, VDSA and DES.

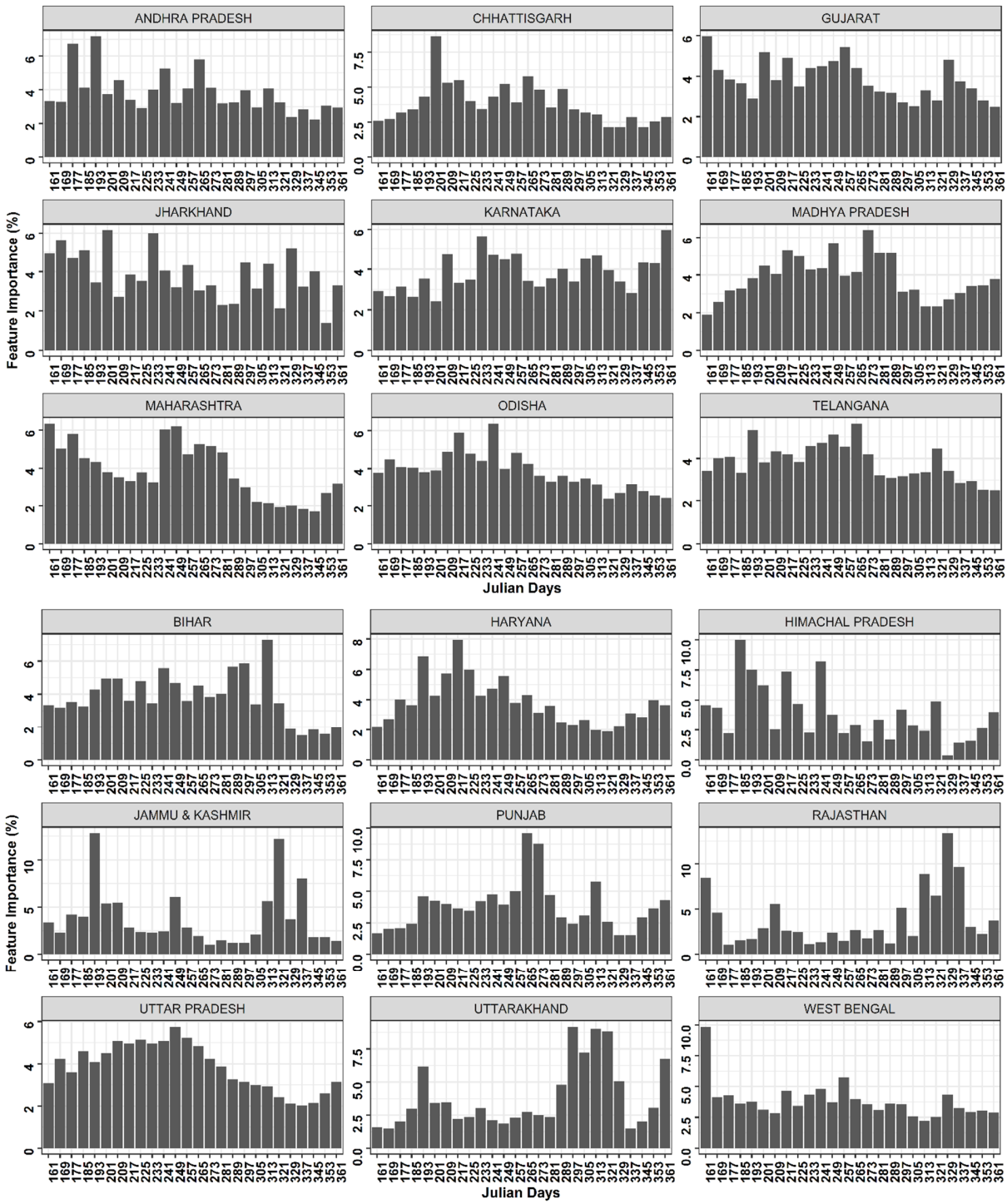

Figure S2. State–wise relative importance (%) of input parameters (time–series LAI values) in GBR.

Table S2. The state–wise aggregated yield (2003 to 2015) of observed and simulated.

State	OBS	SIM
Andhra Pradesh	2.86	2.85
Bihar	1.23	1.30
Chhattisgarh	1.49	1.50
Gujarat	1.77	1.96
Haryana	2.87	2.89
Himachal Pradesh	1.58	1.71
Jammu & Kashmir	1.57	1.79
Jharkhand	1.85	1.83
Karnataka	2.55	2.61
Madhya Pradesh	1.16	1.22
Maharashtra	1.38	1.57
Odisha	1.50	1.52
Punjab	3.81	3.86
Rajasthan	1.87	1.92
Telangana	2.88	2.94
Uttar Pradesh	2.08	2.12
Uttarakhand	1.69	1.96
West Bengal	2.49	2.57

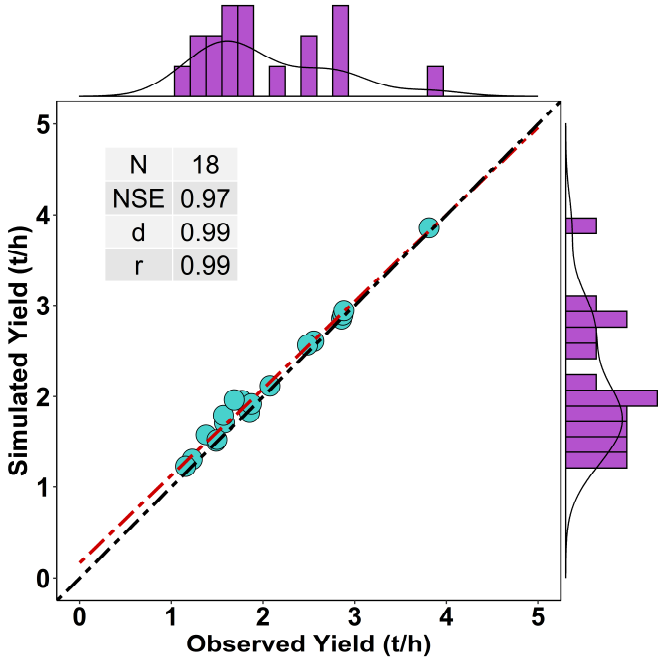

Figure S3: The comparison between state aggregated over the model development years (2003–2015) of observed and simulated yields. The points in the plots are representing states. The red and black lines are representing the trend line and 1:1 line, respectively. The values for the indicators N, NSE, d, and r are representing the number of observations, Nash–Sutcliffe model efficiency coefficient, the index of agreement and the Pearson's correlation coefficient, respectively. The histograms on the plot representing the yield distribution in the observed and simulated yield.

Table S3: The temporal validation between block–level observed and simulated yields. N, d, and r represent observations, the index of agreement, and the correlation coefficient, respectively.

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
N	105	103	104	103	98	96	99	98	87	87	81	83	81
d	0.65	0.72	0.68	0.59	0.62	0.67	0.66	0.72	0.64	0.67	0.58	0.66	0.57
r	0.7	0.81	0.64	0.52	0.7	0.76	0.71	0.74	0.65	0.65	0.67	0.79	0.56

Table S4: The parameters used for analyzing the assessment of the goodness inde**x.**

Statistical parameter	Formula	Reference
The index of agreement (d)	$1 - \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}\right]$	[1]
Nash–Sutcliffe model efficiency coefficient (NSE)	$1 - \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}\right]$	[2]
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert P_i - O_i\rvert$	[3]
Pearson's correlation coefficient (r)	$\frac{\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \bar{P})^2}}$	[4]
Pearson's correlation coefficient of determination (R**2**)	$\left(\frac{\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \bar{P})^2}}\right)^2$	[5]
Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum(P_i - O_i)^2}$	[6]
%Root Mean Square Error (RMSE)	$100 * \sum(O_i)^{-1} * \sqrt{\frac{1}{n}\sum(P_i - O_i)^2}$	[7]

Note: Pi – Time–series simulated yield anomaly; Oi – Time–series observed yield anomaly; n– Number of time series; $\bar{O}$ is the mean of the observed yields

Equation S1: Multiple linear regression model

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_k X_{ik} + \beta_i}$$

Yi = response value for ith individual
X_{ij} = value of the jth independent variable for the ith individual
β_0 = Intercept of regression equation
β_j = coefficient of the jth independent variable
β_i = ith random error component

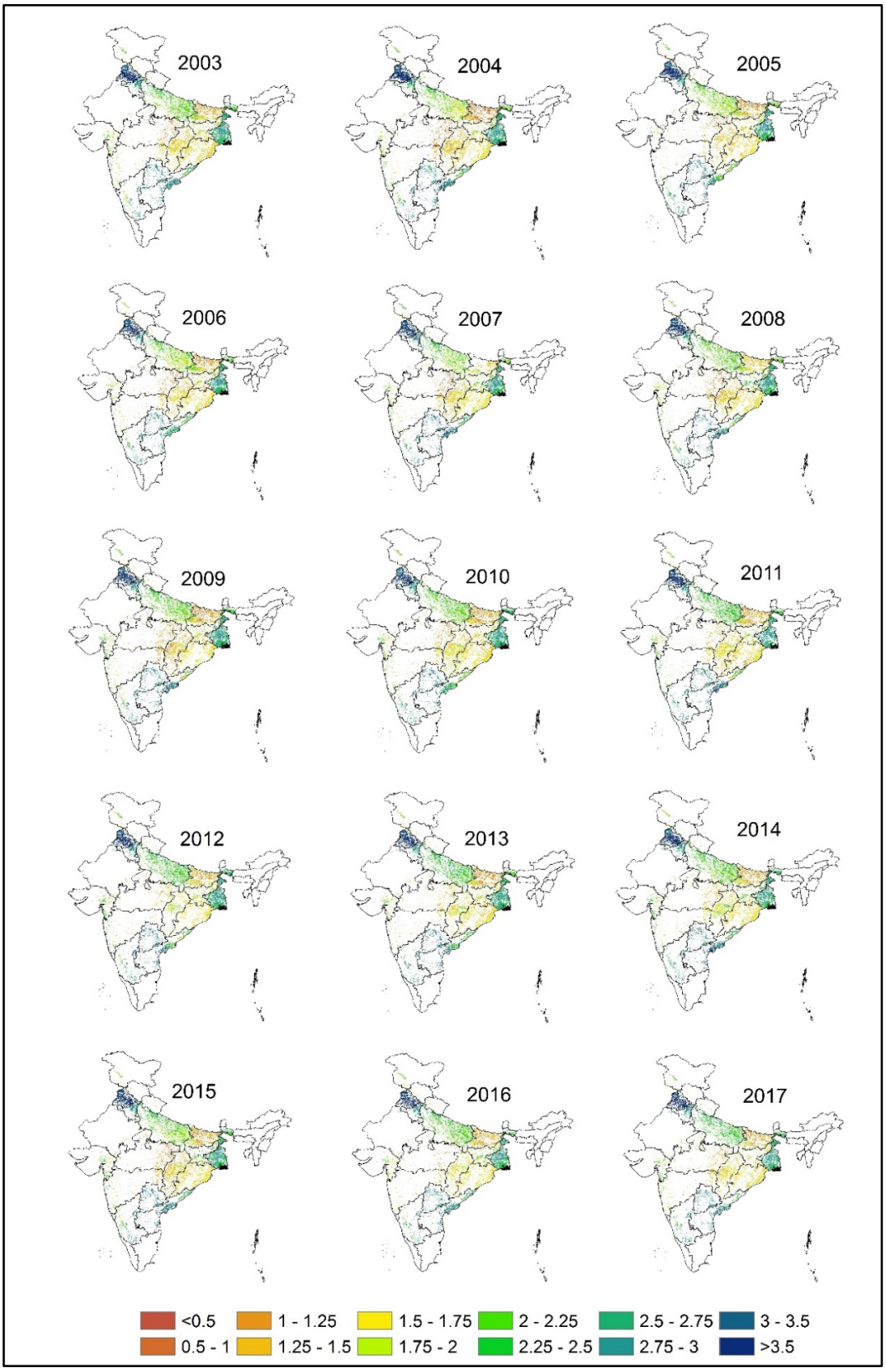

Figure S4: Spatial distribution of simulated yields from the year 2003 to 2017.

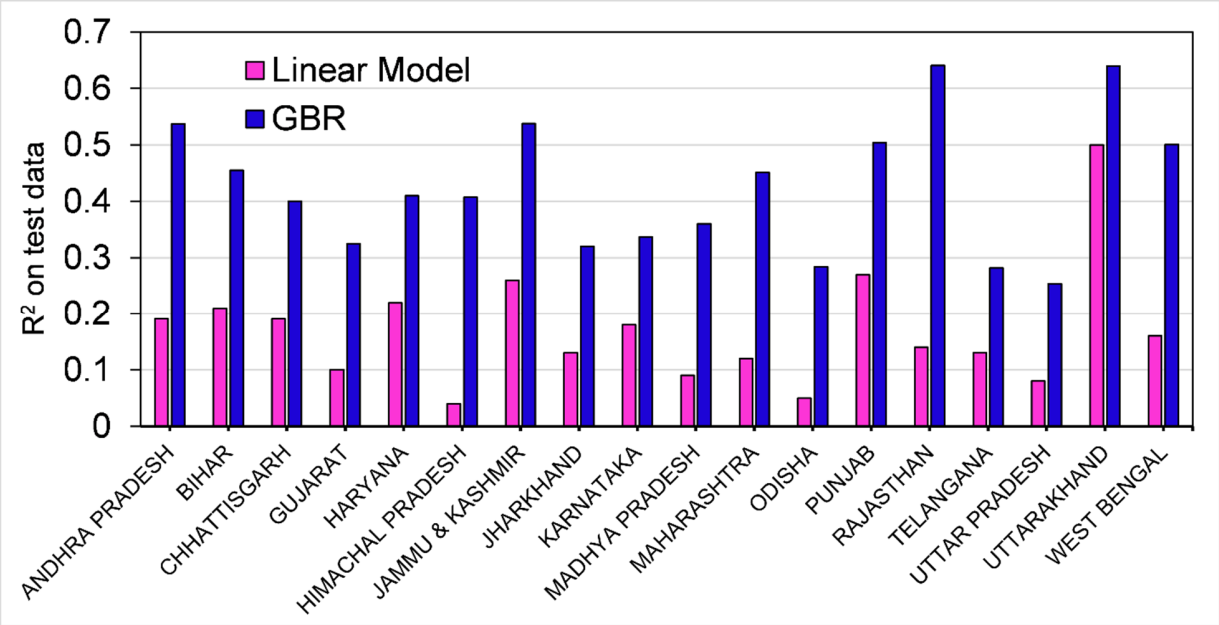

Figure S5: Comparision between GBR and Linear model's (trained with 80% of data) accuracy on test data (validated with 20% of data) for each state.

Table S5: Comparision between GBR and Linear model's (trained with 80% of data) accuracy on test data (validated with 20% of data) for each state with the difference increased in GBR from linear models.

States	Linear Model (R2)	GBR (R2)	% difference
Andhra Pradesh	0.19	0.54	182.7
Bihar	0.21	0.45	116.7
Chhattisgarh	0.19	0.40	110.8
Gujarat	0.1	0.33	225.1
Haryana	0.22	0.41	86.4
Himachal Pradesh	0.04	0.41	917.9
Jammu & Kashmir	0.26	0.54	106.9
Jharkhand	0.13	0.32	146.2
Karnataka	0.18	0.34	87.2
Madhya Pradesh	0.09	0.36	300.2
Maharashtra	0.12	0.45	276.1
Odisha	0.05	0.28	468.2
Punjab	0.27	0.50	86.5
Rajasthan	0.14	0.64	357.4
Telangana	0.13	0.28	116.9
Uttar Pradesh	0.08	0.25	216.9
Uttarakhand	0.5	0.64	28.0
West Bengal	0.16	0.50	213.0

References:

1. Willmott, C.J. On the Validation of Models. *Phys. Geogr.* **1981**, *2*, 184–194.
2. Yuan, F.; Ren, L.–L.; Yu, Z.–B.; Xu, J. River Flow Routing Model Based on Digital Drainage Network. *J. Hydrodyn. Ser. B* **2005**, *17*, 483–488.
3. Taylor, S.L.; Payton, M.E.; Raun, W.R. Relationship between Mean Yield, Coefficient of Variation, Mean Square Error, and Plot Size in Wheat Field Experiments. *Commun. Soil Sci. Plant Anal.* **1999**, *30*, 1439–1447.
4. Sedgwick, P. Pearson's Correlation Coefficient. *Bmj* **2012**, *345*, e4483.
5. Kvålseth, T.O. Cautionary Note about R 2. *Am. Stat.* **1985**, *39*, 279–285.
6. Balkovič, J.; van der Velde, M.; Schmid, E.; Skalský, R.; Khabarov, N.; Obersteiner, M.; Stürmer, B.; Xiong, W. Pan–European Crop Modelling with EPIC: Implementation, up–Scaling and Regional Crop Yield Validation. *Agric. Syst.* **2013**, *120*, 61–75, doi:10.1016/j.agsy.2013.05.008.
7. Smith, M.W.; Then, A.Y.; Wor, C.; Ralph, G.; Pollock, K.H.; Hoenig, J.M. Recommendations for Catch–Curve Analysis. *North Am. J. Fish. Manag.* **2012**, *32*, 956–967, doi:10.1080/02755947.2012.711270.