Does choice of drought index influence estimates of drought-induced rice losses in India?

Francisco Fontes[1,*], Ashley Gorst[2], Charles Palmer[3]

[1] Monitoring and Analyzing Food and Agricultural Policies (MAFAP) program, Agricultural Development Economics Division (ESA), Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, [2] Vivid Economics Ltd., London, UK, and [3] Department of Geography and Environment & Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London, UK

*Corresponding author. Email: frapfontes@gmail.com

ONLINE APPENDIX

Appendix A. Data and variables

The raw data file includes cumulative monthly rainfall data at the district level.

Generating rainfall variables

We start by generating the rainfall variable, which represents cumulative rainfall over the June-September period. A long-term average rainfall measure is then defined for each district. We take the average total cumulative rainfall over the growing season (June-September) for each district over the period 1956-2009.

For a given district, the general formula used is the following:

$$TR_{it} = \sum_{m=1}^{N} R_{mit},$$

where the total rainfall in a given growing season for a given district *i* in a given year *t*, is equal to the sum of the monthly cumulative rainfall over the June-September months (*m* to *M*) included in the growing season. To calculate the long-term average rainfall, we use the following formula:

$$LTAR_i = \frac{1}{54} \sum_{t=1956}^{T=2009} TR_{it},$$

where the long-term average rainfall for a given district *i* is simply calculated as the average total rainfall in that district over the 1956-2009 period.

Generating temperature variables

We opt for a measure of cooling degree days (*CDD*) to capture accumulated heat over the growing season (June-September, in our main specification). This captures the number of degree days above a reference (average) temperature, DTA_i, over a given time period. We use two alternative specifications for generating this variable.

Our first step is to define the average temperature over the growing season for each district between 1956 and 2009. For any given district, *CDD* is estimated as:

$$CDD_{it} = \sum_{m=1}^{M} \sum_{d=1}^{D} (DT_{imd} - DTA_i).$$

Our long-term average *CDD* is then calculated as follows:

$$LTACDD_i = \frac{1}{54} \sum_{t=1956}^{T=2009} CDD_{it},$$

where *d* and *m* represent a given day and month included in the growing season and *D* and *M* respectively represent the total numbers of days in a given month and the total number of months in the growing season; *DT* denotes the average daily temperature in district *i* in day *d* of month *m*; and *DTA* represents the average growing season daily temperature for a given district over the 1956-2009 period. Next we create $LTACDD_i$, which is simply the average cumulative degree days above the mean daily temperature experienced by district *i* over the 1956-2009 period.

Generating drought indices

Crucial to our analysis is the construction of a novel drought index. For our purposes, we develop three drought indices. Below we describe the steps we carry out for each one.

Yu-Babcock index

We denote: total rainfall over the growing season TR_{it}; the mean of total rainfall over the growing season over 1956-2009 $LTAR_i$; and the standard deviation of TR_{it} as $sdTR_i$. We then obtain the standardized variable using the following formula:

$$TR_{it}^{stand} = \frac{TR_{it} - LTAR_i}{sdTR_i}.$$

We proceed analogously for our CDD_{it} measure. Let: CDD_{it} be cumulative cooling degree days above the long-term mean temperature of a district during the growing season; $LTACDD_i$ be long-term average cumulative cooling degree days in the growing season; and $sdCDD_i$ be the standard deviation of CDD_{it}. We compute the standardized variable:

$$CDD_{it}^{stand} = \frac{CDD_{it} - LTACDD_i}{sdCDD_i}.$$

Following this, we use the following to compute the Yu-Babcock index:

$$DI_{it} = \left[-\max\left(0, CDD_{it}^{stand}\right)\right] * \left[\min\left(0, TR_{it}^{stand}\right)\right]. \qquad \text{(A1)}$$

Normalized indices

We start by defining a variable that captures the deviations vis-à-vis the long-term means of CDD and rainfall. Specifically, we calculate the deviations of CDD from the long-term averages by estimating:

$$DCDD_{it} = CDD_{it} - LTACDD_i.$$

Similarly, we calculate deviations of cumulative rainfall by estimating:

$$DTR_{it} = TR_{it} - LTAR_i\ .$$

In contrast to the Yu-Babcock index, for the remaining indices we use a variable normalized between 0 and 1, rather than a standardized value. We construct a variable, MTR_{it}, which is simply the negative of TR_{it} (i.e. $MTR_{it} = -TR_{it}$). The following is estimated to obtain NTR_{it} and $NCDD_{it}$:

$$NTR_{it} = \frac{MTR_{it} - MTR_i^{min}}{MTR_i^{max} - MTR_i^{min}}$$

$$NCDD_{it} = \frac{CDD_{it} - CDD_i^{min}}{CDD_i^{max} - CDD_i^{min}}.$$

We differ from Yu and Babcock (2010) in creating a normalized version of the rainfall and CDD variables such that they vary strictly between 0 and 1, with 1 indicating the most extreme value (the highest CDD and lowest rainfall) and 0 indicating the lowest value. From these two variables, we then create a normalized index, $NRTI_{it}$, which is simply a product of these variables:

$$NRTI_{it} = NTR_{it} * NCDD_{it} .$$

From this, we obtain two additional indices. First, our Type 1 drought index:

$$DI1_{it} = \begin{cases} NRTI_{it} \; if \; DTR_{it} < 0 \;\; and \; DCDD_{it} > 0 \\ 0 \; otherwise \end{cases} .$$

This is equivalent to a normalized version of the Yu-Babcock (2010) index. It only takes a non-zero value for events where rainfall deficiency and CDD are above average.

Second, we create our Type 2 drought index analogously using the following:

$$DI2_{it} = \begin{cases} NRTI_{it} \; if \; DTR_{it} < 0 \;\; and \; DCDD_{it} < 0 \\ 0 \; otherwise \end{cases} .$$

This is the category omitted by Yu and Babcock. It only takes a non-zero value for events where rainfall deficiency is above-average and CDD is below average.

Determining the sample and generating trends

After developing the drought indices, we create a data file which includes only the observations between 1966 and 2009, i.e., our sample period. This choice is driven purely by data availability. Prior to 1966, our dependent variables (production and yields) are missing from the ICRISAT dataset, and hence would have resulted in districts being dropped. Prior to starting our analysis, we also dropped any districts for which at least one observation is missing in order to keep a balanced panel. We then generate district-specific quadratic trends using the following:

$$trend = t - 1965$$

$$trend\_sq = trend^2 ,$$

where *t* denotes the year.

Appendix B. Estimating economic impact

As is made clear in the main text, the cost estimates generated in this paper are based purely on yield losses, without taking into account any potential changes in the cultivated area. Specifically, our cost estimates are derived using a series of seven steps. We detail all the assumptions and steps used throughout and discuss their relative strengths and weaknesses.

Step 1 - Obtain a national estimate of rice prices for each year:

Crop prices. We generate a national weighted average of crop price by year (using the *egen* command and the user-written option *wtmean*), where the weight is determined by area of land under cultivation. As a result, we first generate, for each year, a weighted average of millet prices at the district-level.

We then use 2008 crop prices to estimate prices (and costs) in US$: Rice prices are estimated at 29.947 US$/quintal. These prices are obtained by obtaining the weighted average of rice prices in India for 2008 (in Rupees) and converting this using the averages of the 2008 monthly exchange rates extracted from: http://www.x-rates.com/average/?from=USD&to=INR&amount=1&year=2008.

The results were also computed using nominal yearly prices in Rupees and are available from the authors upon request.

Weaknesses and strengths of the assumptions:

National rice prices. For any given year, there are large differences in prices across districts. It could be argued that prices at the district- or state-level may be more appropriate. However, there are issues with missing price data at the district-level and, to a lesser extent, at the state-level, even for cases where there is a non-zero quantity reported. This is the main reason why we opt for national prices.

Using fixed rice prices in US$. Using a fixed price throughout the sample period implies that the estimates of costs will vary depending on the chosen year since the choice of the year will, by definition, drive both the exchange rate and the price level. Yet, output losses in the early periods are made comparable to losses in later periods since they are given the same value. Using nominal prices could lead to the economic cost of drought artificially increasing over time as nominal prices have trended upwards over the sample period. In any case, we have also performed this exercise using nominal prices in rupees and the results are available from the authors upon request.

Step 2 - Estimate the regression of interest:

We estimate a fixed-effects model per (3) in the main text.

Step 3 - Estimate the yield losses:

After Stata has generated the output for the regression in Step 2, we operationalise the following steps:

- *Step 3.1* – Predict the yield for drought when $DI1_{it} > 0$ or $DI2_{it} > 0$ (i.e., when the given district is drought affected). We do this by using the *levpredict* command following the estimation of the regression before replacing observations not affected by drought with an empty observation. We denote this variable $yhat_d$**.** Note that, to limit potential biases in the estimates of overall costs, we remove districts with implausible predicted yields, which we define as yields below 100 kg/ha and above 5 tonnes/ha). This assumption, however, affects very few observations (less than 0.01% of total events).
- *Step 3.2* – Predict the yield variable under no drought (i.e., when $DI1_{it} = 0$ or $DI2_{it} = 0$). We rename the original variables $DI1original_{it}$ and $DI2original_{it}$, and create two new temporary variables: $DI1temp_{it} = 0$ and $DI2temp_{it} = 0$. We then

use the *levpredict* command to obtain predicted yield and a variable denoted $yhat_{nd}$. The variables $DI1temp_{it}$ and $DI2temp_{it}$ are deleted, and $DI1original_{it}$ and $DI2original_{it}$ are, respectively, renamed $DI1_{it}$ and $DI2_{it}$. We replace $lyhat_{nd}$ with an empty observation for every case where $DI1temp_{it} = 0$ and $DI2temp_{it} = 0$ (non-drought affected case)**.**

- *Step 3.3* – Obtain predicted yield losses by simply subtracting the predicted yield under no drought (Step 3.2) by the actual predicted yield (Step 3.1) for all cases when $DI1_{it} > 0$ or $DI2_{it} > 0$. Formally, we calculate $ylosses = yhat_{nd} - yhat_{d}$**.**
- *Step 3.4* – Obtain predicted yield losses by drought type by simply subtracting the predicted yield under no drought by the actual predicted yield for each type of drought separately. Thus, we estimate: $ylosses_1 = yhat_{nd} - yhat_{d}\ if\ DI1_{it} > 0$; and $ylosses_2 = yhat_{nd} - yhat_{d}\ if\ DI2_{it} > 0$**.** Note that the two types of drought are mutually exclusive (i.e., it is impossible for a district to simultaneously have a Type 1 and a Type 2 drought).

Step 4 - Estimate district-level production losses:

This requires three further steps:

- *Step 4.1* - Convert land area to ha**.** As highlighted in the supporting documentation,[1] the land-use data is in 000's of ha. As a result we simply multiply cereal area by 1,000 to derive the cereal area in ha.
- *Step 4.2* - Convert yield losses to 1,000 t/ha. Currently, our yield losses are in t/ha. We thus convert the yield losses to 1,000 t/ha by dividing $ylosses$ by 1,000.
- *Step 4.3* – Get the total district production losses (in 1,000 t). Obtain the product of the variable obtained in Step 4.1 by that obtained in Step 4.2.

[1] See: http://vdsa.icrisat.ac.in/Include/document/all-apportioned-web-document.pdf .

Step 5 - Estimate the district-level cost of production losses:

To do this we perform two further steps:

- *Step 5.1* – Convert price data to million US$/1,000 t**.** For the results shown in the paper, our price data are in US$ per quintal (as explained in Step 3.1) and our production loss data (estimated in Step 4.3) are in 1,000t. To obtain the price data in million US$ per 1,000 t we divide our price level by 100. Note that a quintal is 100 kg. To convert it into 1,000 t (1,000,000 kg), we multiply the price data by 10,000. However, since we want the data in million US$ rather than US$, we divide this by 1,000,000. Thus, price*10,000/1,000,000 = price/100.
- *Step 5.2* – Obtain total value of production losses. After obtaining prices in million US$/1,000 t, we multiply the variable derived in Step 5.1 by the variable derived in Step 4.3 to obtain the total value of production losses in US$ millions. Note that for our estimates in Rupees, we apply the exact same procedure using yearly nominal prices.

Step 6 - Estimate total yearly production losses:

To obtain this measure in 1,000t, we sum estimated total production losses of each affected district in a given year. We use the *total* function of the *egen* command. Note that the value in table 6 in the main text represents the unweighted average yearly loss.

Step 7 - Estimate total yearly production costs:

To obtain this measure in millions of Rupees, we simply sum the estimated total value of the production losses of each affected district in a given year. We use the *egen* command with the *total* function. Note again that the value in table 6 in the main text represents the unweighted average yearly loss.

Appendix C. Estimating forecasting accuracy

For the results in appendix table A8, we estimate the forecasting accuracy of five different models, namely:

- Model 1: DI1 + DI2 separate
- Model 2: DI1 (normalized Babcock index)
- Model 3: DI1 + DI2 in a unique index
- Model 4: Rainfall index (proportion of rainfall below normal)
- Model 5: CDD index – CDD above long-term average growing-season daily temperature for the district

We define 2000 as the main cut-off point to evaluate the forecasting accuracy of our model. In addition, we also test the sensitivity to the choice of cut-off point by using alternative cut-off points (1990, 1995, and 2004). For each cut-off point, we carry out the following steps:

1. For each model and evaluation period, we estimate the following model (in levels) up to the last year of the evaluation period (e.g., up to 2000), using a fixed effects regression:

$$y_{it} = \alpha_i + \gamma_t + \delta_{i1} * t + \delta_{i2} * t^2 + \beta_{1q} DI_{itq} + \beta_{2q} DI_{itq}^2 + \beta_{3q} DI_{itq} * t + \beta_{4q} DI_{itq}^2 * t + \beta_{5q} DI_{itq} * propirri_{it} + \beta_{6q} DI_{itq}^2 * propirri_{it} + \epsilon_{it}.$$

Note that for models 1-3 we use our drought indices. For model 4, *DI* becomes the rainfall index. For model 5, *DI* becomes the CDD index.

2. Once the relationship is estimated, we predict yields for the six years following the last year included in the regression (i.e., if 2000 is the last year, then we estimate predicted values for 2001-2006) using the coefficients from the model estimated up to the year 2000.

3. We then calculate the difference between the estimated values obtained in step 2 against the observed data.
4. We then calculate the Mean Absolute Error (MAE) by computing the average absolute deviation between the predicted values and the observed values for the evaluation period.
5. We then calculate the Root Mean Squared Error (RMSE) by estimating the average squared-error and then taking the square root of this value.
6. For the False positives (FP) and false negatives (FN), we start by defining a 'normal' yield. We do this by calculating the district-specific median yield in the last 5 years included in the regression in step 1 (i.e., if 2000 is the last year included, a normal yield will be the median yield for the 1995-2000 period).
7. We then define a 'large' deviation from normal as a 10% negative deviation.
8. We then generate a FP dummy variable which takes the value of 1 if our model predicts a yield below 90% of normal (i.e., a yield lower than a 10% negative deviation) and the observed value is above this threshold. The dummy takes a value of 0 otherwise.
9. We then generate a FN dummy variable which takes a value of 1 if our model predicts a yield above 90% of normal when the observed yield was lower than 90% below-normal. The dummy takes a value of 0 otherwise.
10. Finally, we run 100 bootstrap iterations and report the bootstrap standard errors for the RMSE, the MAE, the FN rate and the FP rate.

For the results in appendix table A10, the procedure is identical to the one described for model 1 (which is the *DI1 + DI2* separate model, as before). For the other models, the procedure differs very slightly because the differences in yields are predicted. For the results in appendix table A10, we estimate four alternative models, namely:

- Alt1: Dependent variable in first-differences and *DI1+ DI2* in a separate index. Specifically, we estimate the following model:

$$\Delta y_{it} = \alpha_i + \beta_1 DI1_{it} + \beta_2 DI1_{it}^2 + \beta_3 DI1_{it} * propirri_{it} + \beta_4 DI2_{it} + \beta_5 DI2_{it}^2 + \beta_6 DI2_{it} * propirri_{it} + \beta_7 QI3_{it} + \beta_8 QI4_{it} + \beta_9 propirri_{it} + \beta_{10} propirri_{it}^2 + \epsilon_{it},$$

where Δy_{it} is the first-difference in rice yields (levels), *DI1* is the index for type 1 droughts, *DI2* is the index value for type 2 droughts, $propirri$ denotes the proportion of rice area under irrigation. $QI3$ and *QI4* are the non-drought values analogous to *DI1* and *DI2* (i.e., they represent the index values for years when rainfall was above average and temperature was below- (*QI3*) and above-average (*QI4*), respectively). Note, however, that the inclusion of *QI3* and *QI4* has only a marginal effect on the performance of the forecasting models.

- Alt2: Dependent variable in first-differences and *DI1+ DI2* in a unique index. Specifically, we estimate the following model:

$$\Delta y_{it} = \alpha_i + \beta_1 DI12_{it} + \beta_2 DI12_{it}^2 + \beta_3 DI12_{it} * propirri_{it} + \beta_4 QI3_{it} + \beta_5 QI4_{it} + \beta_6 propirri_{it} + \beta_7 propirri_{it}^2 + \epsilon_{it.}$$

- Alt3: Dependent variable in first-differences and *DI1 + DI2* in separate indices and disaggregated by month (choice of interactions was defined by experimenting with different specifications). The following model is estimated:
    - $\Delta y_{it} =$

$$\alpha_i + \beta_1 DI1june_{it} + \beta_2 DI1july_{it} + \beta_3 DI1july_{it} * propirri_{it} + \beta_4 DI1august_{it} + \beta_5 DI1august_{it} * propirri_{it} + \beta_6 DI1september_{it} + \beta_7 DI1september_{it}^2 + \beta_8 DI1september_{it} * propirri_{it} + \beta_9 DI2june_{it} + \beta_{10} DI2june_{it} * propirri_{it} + \beta_{11} DI2july_{it} + \beta_{12} DI2july_{it} * propirri_{it} + \beta_{13} DI2august_{it} + \beta_{14} DI2september_{it} + \beta_{15} DI2september_{it} * propirri_{it} + \beta_{16} DI2_{it} * propirri_{it} + \beta_{17} QI3june_{it} +$$

$$\beta_{18}\, QI3july_{it} + \beta_{19}\, QI3August_{it} + \beta_{20}\, QI3September_{it} + \beta_{21}\, QI4june_{it} + \beta_{22}\, QI4july + \beta_{23}\, QI4August + \beta_{24}\, QI4september + \beta_{25}\, propirri_{it} + \beta_{26} propirri_{it}^2 + \epsilon_{it}\,.$$

- Alt4: Dependent variable in first-differences and *DI1 + DI2* in a unique index and disaggregated by month (choice of interactions was defined by experimenting with different specifications). The estimated model is given by:
  - $\Delta y_{it} =$

    $$\alpha_i + \beta_1\, DI12june_{it} + \beta_2\, DI12june_{it} * propirri_{it} + \beta_3\, DI12july_{it} + \beta_4\, DI12july_{it} * propirri_{it} + \beta_5\, DI12august_{it} + \beta_6\, DI12august_{it} * propirri_{it} + \beta_7\, DI12september_{it} + \beta_8\, DI12september_{it}^2 + \beta_9\, DI12september_{it} * propirri_{it} + \beta_9\, propirri_{it} + \beta_{10} propirri_{it}^2 + \beta_{17}\, QI3june_{it} + \beta_{18}\, QI3july_{it} + \beta_{19}\, QI3August_{it} + \beta_{20}\, QI3September_{it} + \beta_{21}\, QI4june_{it} + \beta_{22}\, QI4july + \beta_{23}\, QI4August + \beta_{24}\, QI4september + \epsilon_{it}.$$

The main reason why these changes are likely to improve forecasting accuracy is that the original model was not originally intended for forecasting, and addressing our research question required a large number of interactions, which may actually harm forecasting performance. Also, adding intra-annual drought index values is likely to be important as the month in which the drought occurs may have important yield implications.

For each of these models, we define 2000 as the main cut-off point to evaluate the forecasting accuracy of our model. In addition, we also test the sensitivity to the choice of cut-off point by using alternative cut-off points (1990, 1995, and 2004). For each cut-off point, we carry out the following steps:

1. For each model and evaluation period, we estimate the models described above (Alt1-Alt4) up to the last year of the evaluation period (e.g., up to 1999 if 2000 is the evaluation period) using fixed effects.

2. Once the relationship is estimated, we predict yields for the six years following the last year included in the regression (i.e., if 1999 is the last year, then we estimate predicted values for 2000-2005) using the coefficients from the model estimated up to the year 1999. To predict yields, we first predict the difference in yields (which we now denote as $d\widehat{y_{it}}$). For the first period of the forecast, we then compute predicted yield as follows:

$$\widehat{y_{it}} = y_{it-1} + d\widehat{y_{it}}.$$

For all other periods, we add the predicted $d\widehat{y_{it}}$ to the previous year predicted y ($\widehat{y_{it}}$). For example for year t+1:

$$\widehat{y_{it+1}} = \widehat{y_{it}} + d\widehat{y_{it+1}} = y_{it-1} + d\widehat{y_{it}} + d\widehat{y_{it+1}},$$

and so on for all periods.

3. Steps 3-10 are the same as the steps to obtain the estimates in appendix table A8.

Appendix Tables and Figures

Table A1. Correlation coefficients and Spearman correlation coefficients

Correlation coefficients			
	Yu-Babcock	DI1	DI2
Yu-Babcock	1.000		
DI1	0.787	1.000	
DI2	-0.189	-0.313	1.000
Spearman correlation coefficients			
	Yu-Babcock	DI1	DI2
Yu-Babcock	1.000		
DI1	0.994	1.000	
DI2	-0.363	-0.363	1.000

Table A2. R-squared

	Levels	Logs
DI1=DI2		
F-statistic	0.938	2.269
p-value	0.334	0.134
DI1 trends = DI2 trends		
F-statistic	0.846	5.918
p-value	0.359	0.016
All DI1 coeffs= All DI2 coeff		
F-statistic	0.102	3.459
p-value	0.750	0.065

Table A3. F-tests

	Levels			Log-levels		
	FE only	FE + trends	All	FE only	FE + trends	All
R-squared (within)	0.501	0.731	0.744	0.464	0.597	0.632
R-squared (between)	0	0.604	0.792	0	0.064	0.344
R-squared (overall)	0.168	0.645	0.775	0.164	0.22	0.446

Table A4. Marginal elasticities (irrigation) proportion of area irrigated

		Type 1 only	Types 1 and 2 (sep.)	
Variable	Value	Type 1	Type 1	Type 2
		Log-levels		
Irrigated area (%)	0	-0.48***	-0.51***	-0.73***
Irrigated area (%)	20	-0.41***	-0.43***	-0.60***
Irrigated area (%)	40	-0.34***	-0.36***	-0.48***
Irrigated area (%)	60	-0.27***	-0.28***	-0.35***
Irrigated area (%)	80	-0.20***	-0.21***	-0.23***
Irrigated area (%)	100	-0.13***	-0.13***	-0.10

Notes: *** denotes statistical significance at the 1% level. For both types of events, marginal effects are computed at the mean value when affected (DI1=0.493 and DI2=0.207).

Table A5. Marginal elasticities (time)

		Type 1 only	Types 1 and 2 (sep.)	
Variable	Value	Type 1	Type 1	Type 2
		Log-levels		
Year	1966	-0.36***	-0.38***	-0.06
Year	1970	-0.35***	-0.36***	-0.12
Year	1974	-0.33***	-0.35***	-0.17
Year	1978	-0.32***	-0.33***	-0.23**
Year	1982	-0.30***	-0.32***	-0.29***
Year	1986	-0.29***	-0.30***	-0.35***
Year	1990	-0.27***	-0.28***	-0.41***
Year	1994	-0.26***	-0.27***	-0.47***
Year	1998	-0.24***	-0.25***	-0.53***
Year	2002	-0.23***	-0.24***	-0.59***
Year	2006	-0.21***	-0.22***	-0.65***
Year	2010	-0.20***	-0.21***	-0.71***

Notes: ** and *** denote statistical significance at the 5% and 1% level, respectively. For both types of events, marginal effects are computed at the mean value when affected (DI1=0.493 and DI2=0.207).

Table A6. Full sample results index by month

	Levels		Log-levels	
Variables	1	2	3	4
	Linear	Squares	Linear	Squares
DI1 June	-0.077***	-0.103**	-0.070***	-0.169***
	(0.018)	(0.051)	(0.019)	(0.048)
$DI1^2$ June		0.038		0.142**
		(0.074)		(0.064)
DI1 July	-0.057**	-0.124**	-0.084***	-0.207***
	(0.029)	(0.058)	(0.025)	(0.063)
$DI1^2$ July		0.105		0.192**
		(0.083)		(0.087)
DI1 August	-0.183***	-0.233***	-0.095***	-0.209***
	(0.019)	(0.056)	(0.017)	(0.048)
$DI1^2$ August		0.075		0.175**
		(0.082)		(0.078)
DI1 September	-0.215***	-0.134**	-0.304***	-0.07
	(0.020)	(0.052)	(0.026)	(0.057)
$DI1^2$ September		-0.119*		-0.343***
		(0.069)		(0.089)
DI2 June	-0.209***	-0.341**	-0.228***	-0.378**
	(0.039)	(0.132)	(0.052)	(0.172)
$DI2^2$ June		0.5		0.54
		(0.446)		(0.691)
DI2 July	-0.132*	-0.392**	-0.160**	-0.423***
	(0.067)	(0.183)	(0.064)	(0.160)
$DI2^2$ July		1.206		1.067
		(0.862)		(0.706)
DI2 August	-0.345***	-0.452**	-0.234***	-0.561***
	(0.063)	(0.178)	(0.051)	(0.153)
$DI2^2$ August		0.459		1.454**
		(0.788)		(0.657)
DI2 September	-0.104*	-0.320***	-0.223***	-0.439***
	(0.061)	(0.122)	(0.061)	(0.101)
$DI2^2$ September		1.025*		1.179***
		(0.533)		(0.420)
Irrig	0.741***	0.742***	0.477***	0.476***
	(0.153)	(0.155)	(0.130)	(0.132)
$Irrig^2$	-0.255***	-0.255***	-0.152*	-0.150*
	(0.095)	(0.096)	(0.078)	(0.079)
Constant	0.831***	0.833***	-0.295***	-0.282***
	(0.060)	(0.060)	(0.046)	(0.047)

(table continues on next page)

Table A6. Full sample results index by month (*continued*)

Time trends	✓	✓	✓	✓
District fixed effects	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓
N	6996	6996	6996	6996
Number of districts	159	159	159	159
R-squared a	0.733	0.733	0.608	0.611
R-squared w	0.747	0.748	0.629	0.632

Notes: Values in parentheses denote clustered standard errors at the district level. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively. Time trends denote quadratic district-specific trends.

Table A7. F-tests (index by month)

	Levels		Log-levels	
	1	2	3	4
	Linear	Squares	Linear	Squares
F-test type 2 jointly diff from 0				
F-value	37.930	2.248	5.912	5.279
P-value	0.000	0.136	0.016	0.023
F-test type 2 jointly diff from type 1				
F-value	36.842	3.777	7.031	7.690
P-value	0.000	0.054	0.009	0.006

Table A8. Out of sample forecasts results (rice yield)

Model Estimated until 2000. Forecasted period: 2001-2006					
	DI12 (sep.)	DI1 only	DI12 (tog.)	Prop rainfall	CDD index
RMSE	0.68	0.68	**0.68**	0.69	0.69
BSE	0.03	0.03	0.03	0.03	0.03
MAE	0.54	0.54	**0.54**	0.55	0.55
BSE	0.02	0.02	0.02	0.02	0.03
FP	16.7%	16.6%	**16.2%**	16.8%	17.8%
FN	23.9%	24.0%	**23.7%**	24.1%	24.8%
FN + FP	40.6%	40.6%	**39.9%**	40.9%	42.7%

Notes: RMSE stands for Root Mean Squared Error. MAE denotes the Mean Absolute Error. BSE denotes the Bootstrap Standard Errors (100 repetitions). FN is the rate of false negatives. FP is the rate of false positives and FN + FP is the sum of the false positives and false negatives. Numbers in bold denote the models that perform best for a given metric (i.e., lowest RMSE, MAE, FP, FN and FN+FP).

Table A9. Forecasting - Models (for years below the cut-off period)

Variables	DI12 (sep.)	DI1 only	Variables	DI12 (tog.)	Variables	Prop. Rainfall	Variables	CDD index
DI1	-0.758***	-0.599***	DI12	-0.823***	Rainfall	-1.672***	Temperature	-0.004***
	(0.128)	(0.124)		(0.114)		(0.140)		(0.001)
$DI1^2$	0.451**	0.318*	$DI12^2$	0.543***	$Rainfall^2$	1.703***	$Temperature^2$	0.000***
	(0.180)	(0.178)		(0.158)		(0.152)		(0.000)
DI1*time	0.015***	0.013**	DI12*time	0.017***	Rainfall*time	0.022***	Temperature*time	0.000***
	(0.005)	(0.005)		(0.005)		(0.006)		(0.000)
$DI1^2$*time	-0.016**	-0.015*	$DI12^2$*time	-0.018**	$Rainfall^2$*time	-0.024***	$Temperature^2$*time	-0.000***
	(0.008)	(0.008)		(0.007)		(0.006)		(0.000)
DI1*Irrig	0.204	0.177	DI12*Irrig	0.169	Rainfall*Irrig	0.521***	Temperature*Irrig	-0.001
	(0.134)	(0.133)		(0.117)		(0.169)		(0.001)
$DI1^2$*Irrig	-0.138	-0.101	$DI12^2$*Irrig	-0.112	$Rainfall^2$*Irrig	-0.513***	$Temperature^2$*Irrig	0
	(0.180)	(0.181)		(0.152)		(0.184)		(0.000)
DI2	-0.777**							
	(0.393)							
$DI2^2$	0.006							
	(1.404)							
DI2*time	-0.008							
	(0.017)							
$DI2^2$*time	0.092							
	(0.062)							

(table continues on next page)

Table A9. Forecasting - Models (for years below the cut-off period) (*continued*)

Variables	DI12 (sep.)	DI1 only	Variables	DI12 (tog.)	Variables	Prop. Rainfall	Variables	CDD index
DI2*Irrig	1.010**							
	(0.431)							
$DI2^2$*Irrig	-3.704**							
	(1.551)							
Irrig	0.561***	0.489***	Irrig	0.557***	Irrig	0.488***	Irrig	0.472**
	(0.161)	(0.167)		(0.163)		(0.159)		(0.183)
$Irrig^2$	-0.217*	-0.184	$Irrig^2$	-0.210*	$Irrig^2$	-0.201*	$Irrig^2$	-0.157
	(0.114)	(0.117)		(0.116)		(0.114)		(0.116)
Constant	0.815***	0.796***	Constant	0.816***	Constant	0.838***	Constant	1.015***
	(0.040)	(0.042)		(0.040)		(0.038)		(0.058)
Time trends	✓	✓		✓		✓		✓
District fixed effects	✓	✓		✓		✓		✓
Year fixed effects								
N	5565	5565		5565		5565		5565
Number of districts	159	159		159		159		159
R-squared a	0.712	0.705		0.711		0.717		0.697
R-squared w	0.729	0.722		0.728		0.734		0.715
R-squared b	0.744	0.736		0.742		0.727		0.779
R-squared o	0.739	0.731		0.737		0.729		0.757

Notes: Values in parentheses denote clustered standard errors at the district level. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively. Time trends denote quadratic district-specific trends. Rainfall refers to the proportion of rainfall relative to the long-term average rainfall for all observations where the deviation is negative.

Table A10. Performance of out of sample forecasts results (alternative models, average)

Average performance over 4 cut-offs periods (1990, 1995, 2000, 2005)					
	DI12 (sep.)	Alt1	Alt2	Alt3	Alt4
RMSE	0.665	0.462	0.461	0.466	**0.460**
MAE	0.531	0.348	0.346	0.347	**0.343**
FP	0.188	0.091	0.091	**0.086**	0.089
FN	0.169	0.161	0.160	**0.153**	0.156
FN+FP	0.357	0.251	0.252	**0.240**	0.245

Notes: RMSE stands for Root Mean Squared Error. MAE denotes the Mean Absolute Error. BSE denotes the Bootstrap Standard Errors (100 repetitions). FN is the rate of false negatives. FP is the rate of false positives and FN + FP is the sum of the false positives and false negatives. Numbers in bold denote the models that perform best for a given metric (i.e., lowest RMSE, MAE, FP, FN and FN+FP). The reported forecasting accuracy indicators are the average of the metrics estimated for four different cut-off points (1990, 1995, 2000, 2004). For each of these cut-off points, the models were estimated until *t*-1 and then yields were forecasted for *t* until *t*+5. See explanation in Appendix C.

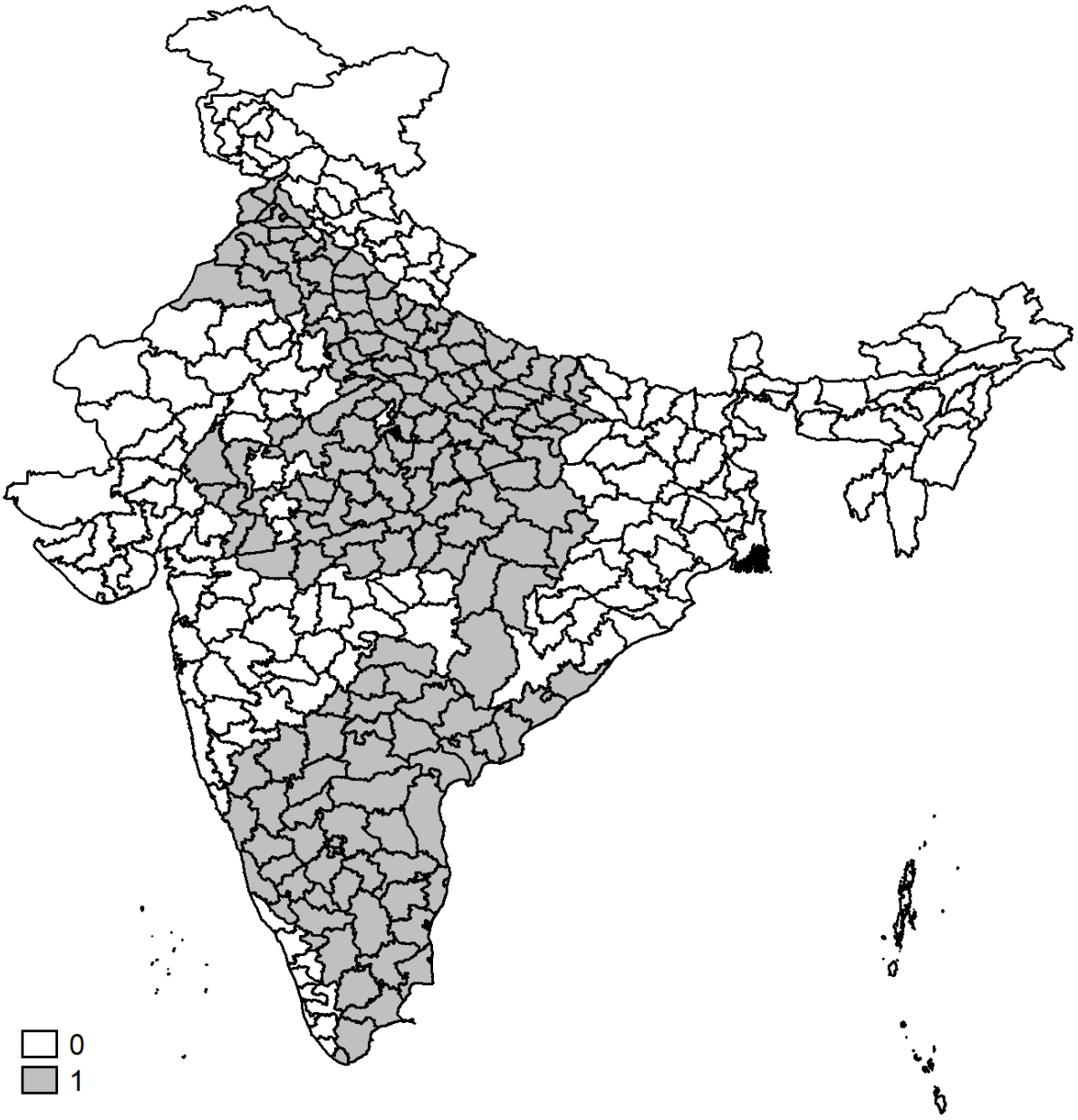

Figure A1. Districts used in the sample.

Notes: Districts in grey are those included in the final sample used for the estimation.