Linear Prediction of Monsoon Rainfall for Indian Subdivisions

Timothy DelSole* and J. Shukla

George Mason University, Fairfax, VA and

Center for Ocean-Land-Atmosphere Studies, Calverton, MD

January 17, 2006

*Corresponding Author Address: Timothy DelSole, Center for Ocean-Land-Atmosphere Studies, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705-3106. Email address: delsole@cola.iges.org

Abstract

The linear prediction of monsoon rainfall for 29 Indian subdivisions and all India is considered. The predictors were selected by computing the regression model for every possible combination of predictors from a pool of ten predictors, and then selecting the combination that leads to the smallest mean square error for the cross validated forecasts. For all India and for most subdivisions, the results clearly demonstrate that prediction models with three or fewer predictors perform much better than models with many more predictors. The optimum prediction model for some subdivisions can include as many as six predictors. About half of the subdivisions appear to be predictable, in the sense that the optimum linear prediction model for these regions have statistically significant skill. The tendency of Darwin pressure emerges as an important predictor, consistent with previous studies, while the index of the quasi-biennial oscillation appears to be relatively useless. These results raise questions about the prediction models currently in use at the Indian Meteorological Department with 8-10 predictors and 17-21 fitted parameters.

1. Introduction

In 1988, the India Meteorological Department (IMD) adopted the so-called 16-parameter power regression model of Gowariker et al. (1989) as part of their official forecasts of Indian monsoon rainfall. Although this model employs 16 observed predictors, the model itself contains 49 independent parameters (DelSole and Shukla 2002). Furthermore, the 49 parameters of the power regression model were estimated from 37 years of training data, suggesting the possibility of overfitting. DelSole and Shukla (2002) further suggested that the "reasonably accurate" predictions by the power regression model were only apparent because the skill of the forecasts were verified during the period 1989-2000. As noted by DelSole and Shukla (2002), this period was unusual in the sense that predictions based on the previous climatological mean had unusually high skill. Indeed, only one year during this period had rainfall anomalies exceeding one standard deviation, an event which had not occurred in any twelve year period since the 1930's (according to the data of Patharsarthy et al. (1995), available from http://www.tropmet.res.in/data.html). Furthermore, the one year in which the rainfall did exceed a standard deviation, namely the 1994 floods, was predicted by the IMD to be a moderate drought. The next significant event was the drought of 2002, which was predicted by the IMD to have slightly above normal rainfall. In 2003, the model was modified to make use of only 10 predictors, but this model still contained at least 21 parameters that were estimated from 38 years of observations (Rajeevan et al. 2004), again suggesting the possibility of overfitting. These statistical and empirical considerations raise questions about the predictive usefulness of the power regression model, and the use of a relatively large number of predictors. Given the tremendous societal importance of the Indian monsoon and associated predictions, these questions deserve serious attention by the scientific community.

The purpose of this paper is to provide strong evidence that statistical prediction models of Indian monsoon rainfall with many parameters have less skill than models with fewer parameters, e.g., a half dozen. This evidence comes not only from predictions of total Indian monsoon rainfall, but also for predictions of rainfall in individual subdivisions. Furthermore, this paper examines the relative importance of predictors in different subdivisions and shows that the tendency of Darwin pressure emerges as a major predictor, consistent with previous studies, and that the quasi-biennial oscillation has questionable merit as a predictor of recent monsoon rainfall.

2. Data

The time series used in this study (and their sources, in parentheses) are yearly values in the period 1951-2002 and are defined as follows:

a. Indian monsoon rainfall in 29 subdivisions averaged over June-September, estimated from observations at 306 land stations uniformly distributed over India (Parthasarathy et al. 1995). The 29 subdivisions and their fractional areal coverage weight are tabulated in table 1.

b. **dtend**: Darwin sea-level pressure tendency: March-April-May average minus December-January-February average. (NCEP, ftp.ncep.noaa.gov/pub/cpc/wd52dg/data/indices/)

c. **nino34mam**: NINO3.4 (Pacific surface temperature over 170°W-120°W, 5°S-5°N), March-April-May average. (Hadley Center, Rayner et al. 2003, http://hadobs.metoffice.com/hadisst/)

d. **naojf**: NAO (sea-level pressure difference between Gibraltar and Stykkisholmur, Iceland) January-February mean. (University of East Anglia; available from www.cru.uea.ac.uk)

e. **naoam**: NAO (sea-level pressure difference between Gibraltar and Stykkisholmur, Iceland): April-May mean. (same as d)

f. **qbojfm**: QBO index, January-February-March average (available from

http://www.cdc.noaa.gov/ClimateIndices/Analysis/#QBO).

g. **wpacmam**: SST averaged in the western Pacific region 120°E-160°E, 5°S-5°N, March-April-May average (same as c).

h. **eindmam**: SST averaged in the eastern Indian Ocean region 70°E-100°E, 5°S-5°N, March-April-May average (same as c).

i. **arabmam**: SST averaged in Arabian Sea region 50°E-70°E, 5°N-15°N, March-April-May average (same as c).

j. **teurodjf**: Eurasian surface temperature (30°E-50°E, 60°N-70°N), December-January-February average. (Jones and Moberg 2003; available from http://www.cru.uea.ac.uk/cru/data/temperature/).

k. **tindiamam**: Indian surface temperature (55°E-75°E, 25°N-35°N), March-April-May average (same as j).

The above time series differ from those used in DelSole and Shukla (2002) in several ways. First, the predictor "ridge," equal to the latitude of the 500hPa ridge at 75°E during April, has been dropped, since its value exceeding four standard deviations of the 1951-1996 climatological distribution after 1997. Second, the tendency of NINO3.4, and the April average minus January average Darwin pressure, have been dropped because they are highly correlated with (b) and (c). Third, the following predictors have been added: an index of the quasi-biennial oscillation, and indices of the sea surface temperatures in the Indian ocean, Arabian Sea, and western Pacific. These latter predictors have been added as proxies of predictors used by the Indian Meteorological Department (IMD) in their most recent forecasts (Rajeevan et al. 2004). We would have preferred to use the same predictors as used by the IMD, but this data is not publicly available.

3. Methodology

The prediction model considered in this paper is the linear regression model

$$y(t) = \beta_1 x_1(t) + \beta_2 x_2(t) + \ldots + \beta_K x_K(t) + e(t), \quad (1)$$

where y is the predictand, here taken to be monsoonal rainfall, $x_1, x_2, \ldots, x_K$ are K known predictors, e is an error term, and $\beta_1, \beta_2, \ldots, \beta_K$ are unknown regression parameters to be estimated from data. A constant term is included by introducing an additional "predictor" whose value is always unity. If the above regression equation is written in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad (2)$$

then the least squares estimates of the regression parameters are

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{y} \quad (3)$$

(von Storch and Zwiers 1999, pg. 159). The least squares prediction of a new set of predictands $\boldsymbol{y'}$ based on a new set of predictors $\boldsymbol{X'}$ is given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}'\hat{\boldsymbol{\beta}}. \quad (4)$$

We estimate the skill of the least squares model by "leave-one-out" cross validation. In this procedure, the time series of length N is split into two parts, one of which is of length N - 1 and used to construct least squares estimates of β from (3), called the "training data," and the other is of length one and used to verify the least squares prediction from (4), called the "verification data." The difference between the resulting prediction and the verification of y is called the cross-validated

prediction. This procedure is repeated, leaving out a different predictand in turn, until all predictands have been used exactly once as verification. The skill of the model is then measured by the root mean square (rms) of the cross validated errors. See Michaelson (1987) for more details of the technique.

It should be recognized that the use of cross validation to select predictors *and* to measure skill leads to biased estimates of skill (DelSole and Shukla 2002). Many regression procedures mitigate this problem by adopting a model selection criterion that balances the apparent increase in skill, as measured in the training data, against the increase in the number of predictors; for example, Akaike's Information Criterion (Burnham and Anderson 2002) or the method of DelSole and Shukla (2002). Unfortunately, conclusions derived from the selected models are subject to the criticism that they depend on the criterion used to select the models. Since our goal is primarily to shed light on the optimum number of predictors in monsoon prediction models, we have avoided the use of additional selection criteria and simply show the variation of cross validated skill for all possible models. However, the bias inherent in our procedure should be kept in mind when interpreting the results.

The "explained variance" EV of a prediction model is defined in this paper as

$$EV = 1 - \frac{\langle (y - \hat{y})^2 \rangle}{\langle (y - \langle y \rangle)^2 \rangle}, \qquad \textbf{(5)}$$

where brackets indicate a time average. The numerator is the rms error while the denominator is the sample variance. EV vanishes when all forecasts equal the sample mean, equals unity for a perfect prediction, and is negative for a forecast "worse" than a prediction based on the climatological mean.

In this paper, we consider all possible subsets of the available predictors. Since there are $K = 10$ total predictors, there are $2^{10} = 1024$ distinct linear regression models (including the constant $x(t) =$

1). A useful method for organizing the computations is to represent each number from 0 to 2^K-1 in its binary representation, assign a one-to-one correspondence between each predictor and each digit of the binary number, and then to include only those predictors which have a "one" in the corresponding digit. In this way, a single algorithmic loop can systematically search every possible combination, and all results can be archived by identifying each unique model with its corresponding number.

4. Results

The root mean square error of the cross validated predictions of all possible linear regression models for total JJAS Indian monsoon rainfall is shown in fig. 1. The dashed line indicates the standard deviation of total Indian monsoon rainfall for the period 1951-2002. The figure reveals that the best prediction model has three physical predictors and explains about 16% of the variance. The associated optimal predictors are *dtend*, *naoam*, and *tindiamam*. As discussed in DelSole and Shukla (2002), the fraction of explained variance and the precise optimal predictors are sensitive to the choice of period and set of predictors. The skill of the best prediction model with two predictors– which turn out to be *dtend* and *naoam*– is nearly indistinguishable from that of the optimum prediction model with three predictors, suggesting that the predictor *tindiamam* adds relatively little predictive skill. The fact that the two predictors *dtend* and *naoam* arise as prominent predictors here is consistent with the findings of DelSole and Shukla (2002, 2006).

We repeated the above procedure but for predicting JJAS rainfall in each of the 29 subdivisions comprising the total monsoon rainfall. The optimal linear model in each subdivision is tabulated in table 1. We see that the optimum number of predictors lies in the range 0-6, with more than two-thirds of the subdivisions having 3 or fewer predictors. The correlation skill of the optimal models is also tabulated. For reference, the 1% significance level for the correlation coefficient with

50 samples is 0.32; correlations exceeding this have been indicated in bold. Thus, table 1 indicates that a little over half of the subdivisions have statistically significant skill. The spatial structure of the correlation skill is indicated in fig. 2 and suggests that the western half of India is "more predictable" than the eastern half.

Maps of the subdivisions in which a given predictor is selected in the optimum linear prediction model are shown in fig. 3. By far the most commonly selected predictor is *dtrend*, an index of Darwin pressure tendency, which is selected in 21 of the 29 subdivisions. This result is consistent with the finding of Shukla and Paolino (1983) that the tendency of Darwin is a useful predictor of Indian monsoon rainfall. Next most common is *teurodjf*, an index of European surface temperature, which is selected in 11 out of 29 subdivisions, but most of these subdivisions overlap with the subdivisions in which *teurodjf* also is selected and therefore seems to be less useful. The predictor *tindiamam*, an index of Indian surface temperature in spring, is selected in 9 out of 29 subdivisions, but these subdivisions extend further eastward into the "unpredictable regions." The latter results may explain why *tindiamam* was selected as a useful predictor for total monsoon rainfall, namely because it is a useful predictor in the eastern region of India in which *dtrend* does not appear to be as useful. Interestingly, *qbojfm*, an index of the quasi-biennial oscillation, is not selected in any subdivision, suggesting that this time series is not as useful as other predictors.

The above approach is not necessarily a good basis for distinguishing predictable regions, because the skill of the optimum prediction model is often not statistically different from other models with fewer predictors. Two regions that appear to be influenced by different sets of predictors according to the above criterion, may in fact be influenced predominantly by the same set of predictors because the additional predictors add so little skill that the difference is not statistically

distinguishable. Therefore, we do not advocate the above approach as a basis for defining "homogeneous regions."

It turns out that predicting each subdivision separately and summing the results improves, albeit marginally, the skill of the total Indian monsoon forecast: the correlation skill of the aggregated predictions is 0.49, compared to 0.41 for the skill of predicting total rainfall directly.

5. Discussion

The above results, especially fig. 1, show very clearly that a linear prediction based on ten predictors is demonstrably worse than many of the models based on fewer predictors. While Rajeevan et al. (2004) explicitly recognized that the optimum number of predictors is a matter of controversy, they stated that "[our] own assessment is that 8 to 10 predictors are required for . . . limiting the root mean square error of the results over the independent period to a minimum." We are baffled by this statement because the power regression model of Rajeevan et al. (2004), of the form

$$y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k^{p_k}, \qquad (6)$$

contains $2K+1$ regression parameters for K predictors, which implies that their models contain 17 to 21 regression parameters. We are unable to understand how the power regression model can achieve a minimum error variance in independent data using 17-21 parameters, when the regression models considered in this paper, which are special cases of the power regression model, achieve a minimum usually with three or fewer parameters. While it is true that the power regression model is nonlinear, nonlinearity often increases the difficulty of estimating and justifying the relevance of predictors, rather than the reverse. Similarly, while it is true that the predictors used by the IMD are not available

to us, the predictors used in this paper are similar to those in Rajeevan et al. (2004). We are not aware of any evidence, such as a plot analogous to fig. 1, that shows the skill of the optimal power regression model for all possible combination of predictors, and demonstrates that models with 17- 21 parameters gives the best forecasts. It would be quite easy for us to produce a figure analogous to fig. 1 for IMD predictions, but IMD, according to Rajeevan (personal communication), cannot make their data available to us.

The results of this paper also shed some light on the spatial structure of Indian monsoon rainfall predictability. Western India appears to be the most predictable region, with *dtrend* and *tindiamam*, which are indices of Darwin pressure tendency and European surface temperature, being the dominant predictors in this region. The importance of *dtrend* as a predictor of monsoon rainfall is consistent with previous studies, most notably Shukla and Paolino (1983), which indicate a relation between monsoon rainfall and the tendency of Darwin prior to the monsoon season. The importance of *teurodjf* as a predictor also is consistent with previous studies which have suggested a link between monsoon rainfall and snowfall in Europe (Bamzai and Shukla 1999). Only limited regions in eastern India appear to be predictable, with the predictor *naojf* (an index of the North Atlantic Oscillation) being the most common predictor in this region. Although this result indicates a connection between monsoon rainfall and variability in the North Atlantic Oscillation, the underlying mechanism is unclear. Finally, the above results suggest that indices of the quasi-biennial oscillation, currently used by the IMD (Rajeevan et al. 2004), are relatively useless predictors of monsoon rainfall.

6. Acknowledgements

This research was supported by the National Science Foundation (ATM0332910), National Aeronautics and Space Administration (NNG04GG46G), and the National Oceanographic and Atmospheric Administration (NA04OAR4310034).

7. References

Bamzai, A. S., and J. Shukla, 1999: Relation between Eurasian snow cover, snow depth, and the Indian summer monsoon: An observational study. *J. Climate*, **12**, 3117–3132.

Burnham, K. P., and D. R. Anderson, 2002: *Model Selection and Multimodel Selection: A Practical Information Theoretic Approach*. 2d ed. Springer-Verlag, 488 pp.

DelSole, T., and J. Shukla, 2002: Linear prediction of Indian monsoon rainfall. *J. Climate*., **15**, 3645–3658.

DelSole, T., and J. Shukla, 2006: Correction note on "Linear prediction of Indian monsoon rainfall." *J. Climate*. Submitted.

Gowariker, V., V. Thapliyal, R. P. Sarker, G. S. Mandal, and D. R. Sikka, 1989: Parametric and power regression models: New approach to long range forecasting of monsoon rainfall in India. *Mausam*, **40**, 115-122.

Jones, P.D. and Moberg, A., 2003: Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Climate*, **16**, 206-223.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Appl. Meteor*., **26**, 1589–1600.

Parthasarathy, B., A. A. Munot, and D. R. Kothawale, 1995: Monthly and seasonal rainfall seies for all India, homogeneous regions and meteorological subdivisions: 1871-1994. Research Report No. RR-065, 113pp. [Available from Indian Institute of Tropical Meteorology, Homi Bhabha Road, Pune 411008, India].

Rajeevan, M., D. S. Pai, S. K. Dikshit, and R. R. Kelkar, 2004: IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003.

Current Science, **86**, 422-431.

Rayner, N.A., D.E. Parker, E.B. Horton, C.K. Folland, L.V. Alexander, D.P. Rowell, E.C. Kent, A. Kaplan, 2003: Global analyses of SST, sea ice and night marine air temperature since the late nineteenth century, *J. Geophys. Res*., **108**, doi:10.1029/2002JD002670.

Shukla, J. and D. A. Paolino, 1983: The Southern Oscillation and long-range forecasting of the summer monsoon rainfall over India. *Mon. Wea. Rev*., **111**, 1830-1837.

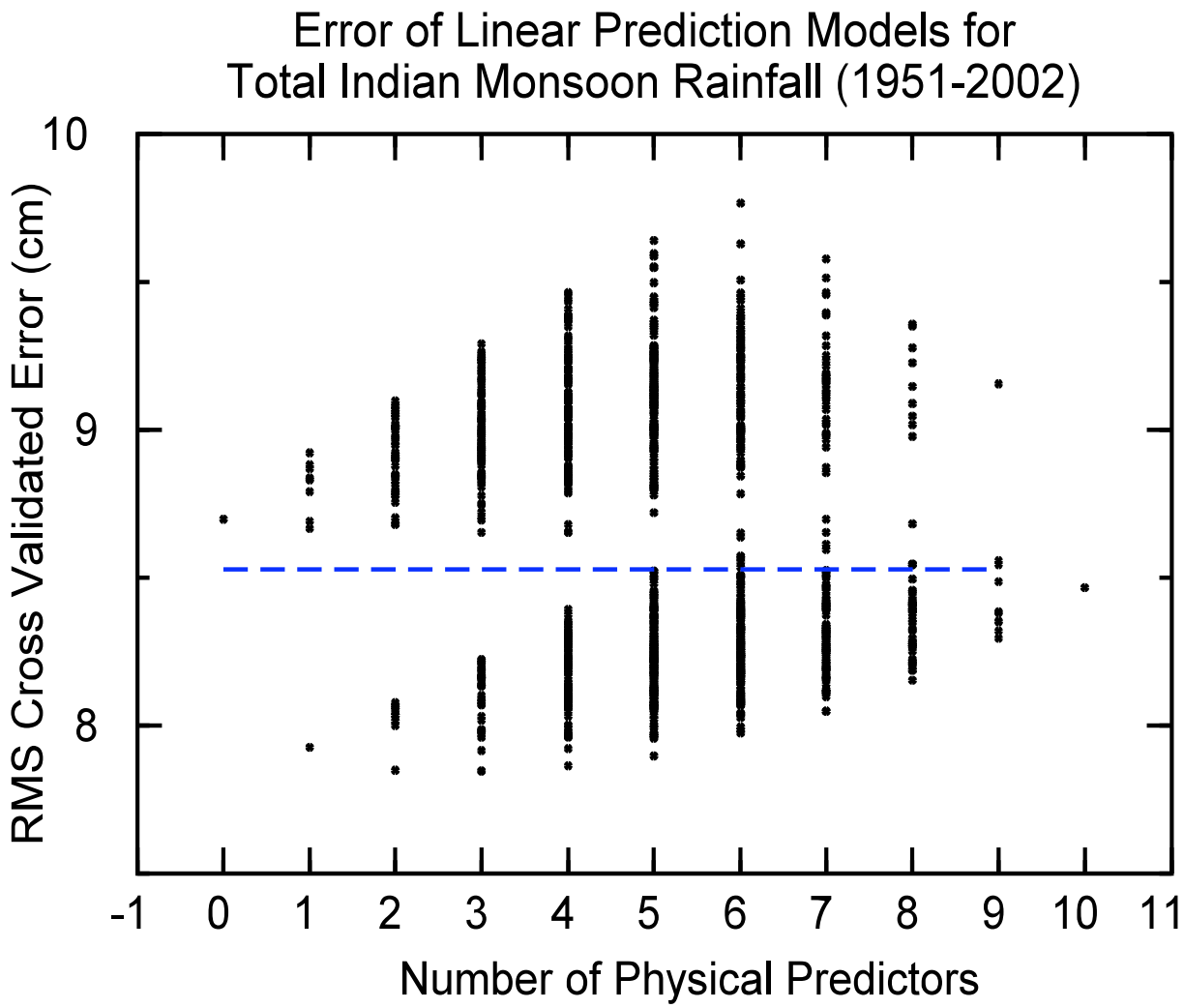

Figure 1: The root mean square error (in cm) of cross validated forecasts of total JJAS Indian monsoon rainfall in the period 1951-2002, based on all possible linear regression models formed from the ten predictors listed in sec. 2, as a function of the number of physical predictors. The case of "zero" physical predictors corresponds to a constant term. (The total number of predictors is one plus the number of physical predictors.) The dashed line indicates the standard deviation of total Indian monsoon rainfall during this period.

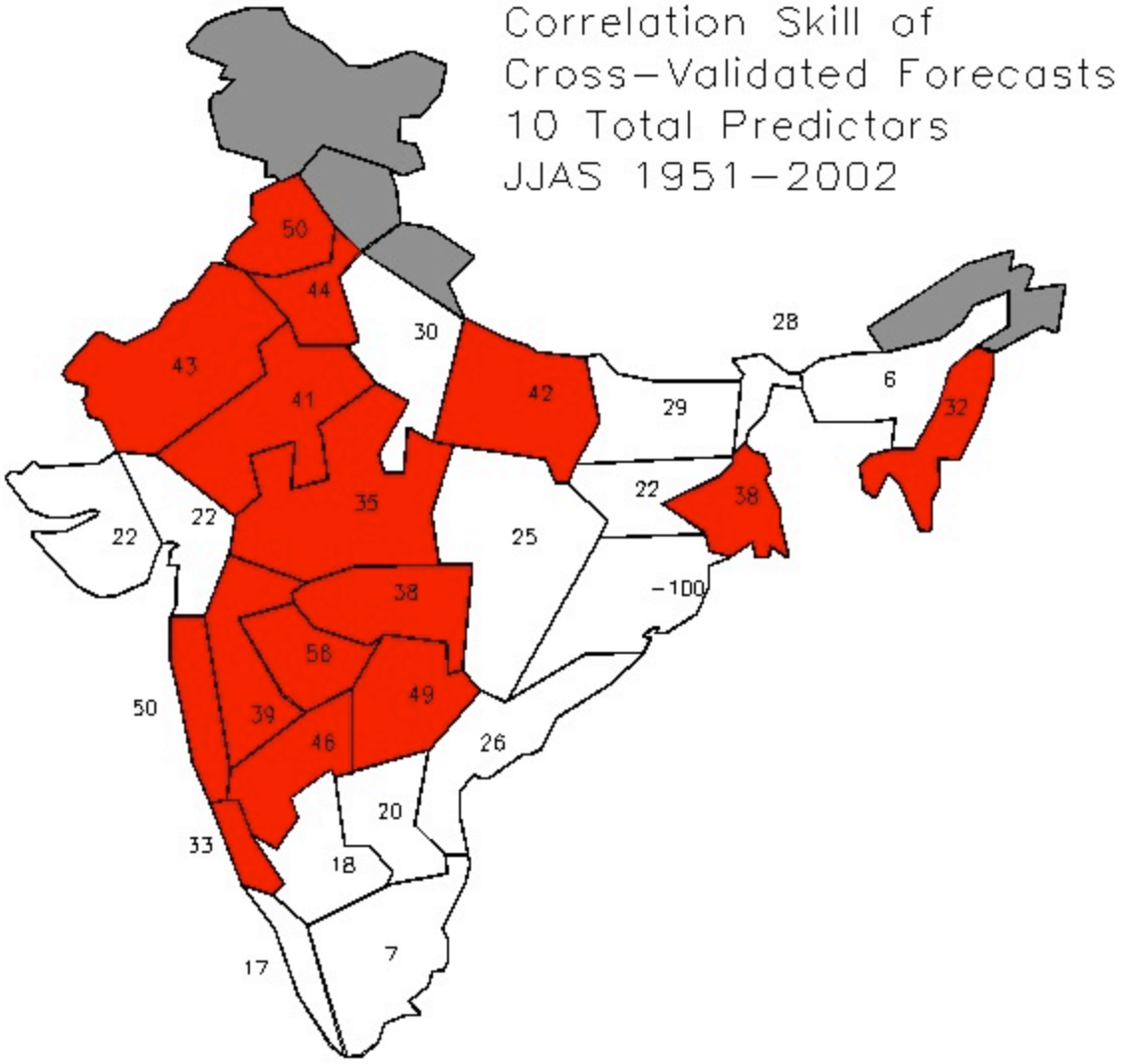

Figure 2: The correlation (in percent) between cross validated forecasts and verification of the optimum linear prediction model for JJAS monsoon rainfall in each subdivision during the period 1951-2002. The optimum prediction model is defined as the model with the minimum root mean square cross validated error out of all possible models formed by the ten predictors defined in sec. 2. The optimum model is chosen in each subdivision independently of the others. Subdivisions in which the correlation exceeds the 1% significance level are shaded in red. Subdivisions which were not part of the analysis are shaded in grey.

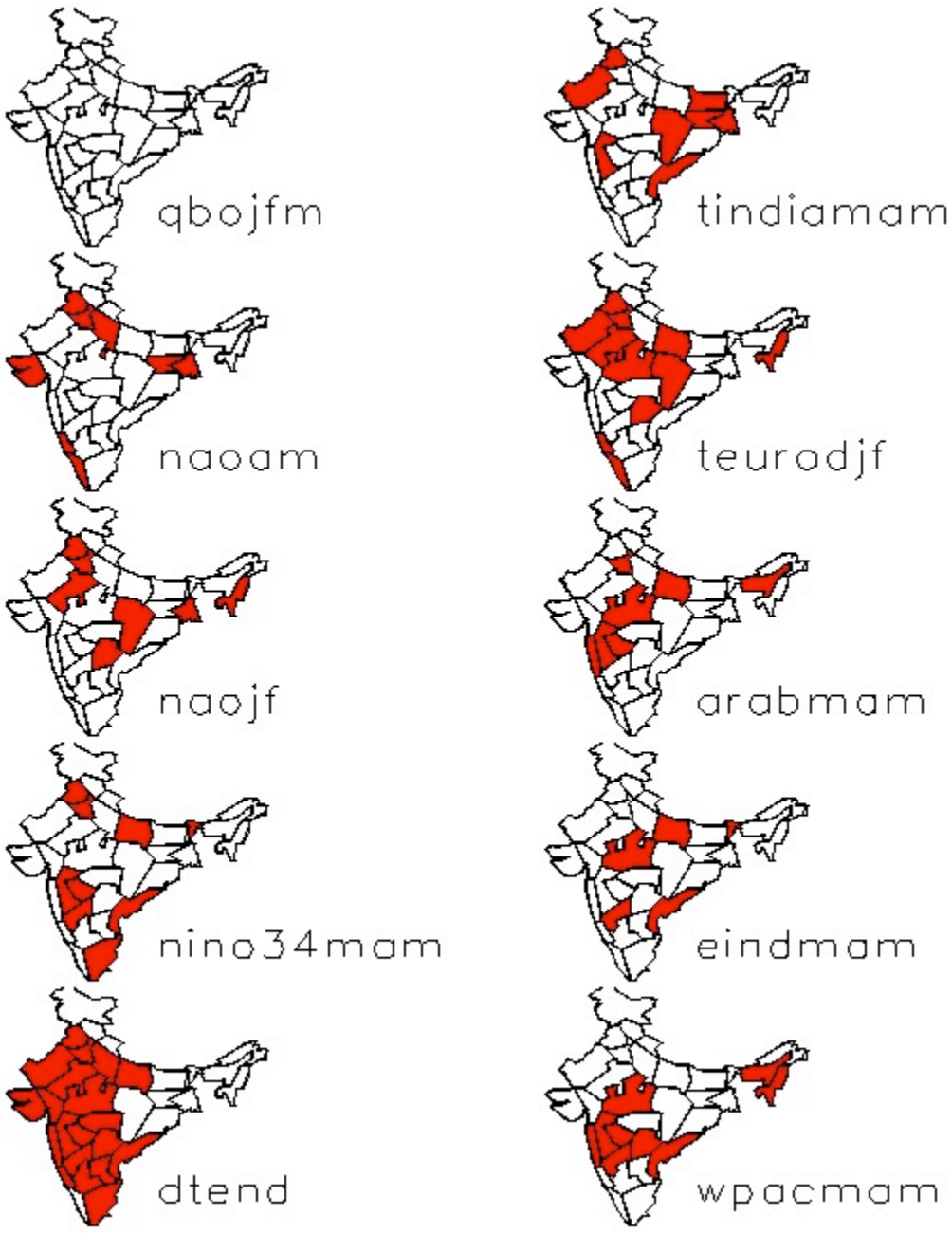

Figure 3: Subdivisions in which the indicated predictors were chosen as part of the optimum linear prediction model of JJAS monsoon rainfall in that subdivision for the period 1951-2002.

subdiv	subdiv	rmse(cm)	stdv(cm)	corr	EV(%)	mean(cm)	weight(%)
NORTH ASSAM	1	18.0	17.7	0.06	-3.4	143.0	2.0
SOUTH ASSAM	**2**	**19.5**	**20.4**	**0.32**	**8.6**	**140.0**	**4.3**
SUB-HIMA. W. BENGAL	3	29.7	30.7	0.28	6.4	198.0	0.8
GANGETIC W. BENGAL	**4**	**17.2**	**18.4**	**0.38**	**12.6**	**117.0**	**2.3**
ORISSA	5	16.6	16.3	-1.00	-3.7	113.0	5.4
BIHAR PLATEAU	6	17.3	17.5	0.22	2.3	108.0	2.8
BIHAR PLAINS	7	18.2	18.9	0.29	7.3	101.0	3.3
EAST UTTAR PRADESH	**8**	**16.7**	**18.2**	**0.42**	**15.8**	**89.4**	**5.1**
WEST U.P. PLAINS	9	15.0	15.6	0.30	7.5	76.7	3.4
HARYANA	**10**	**12.8**	**13.9**	**0.44**	**15.2**	**49.2**	**1.6**
PUNJAB	**11**	**15.3**	**17.3**	**0.50**	**21.8**	**54.8**	**1.7**
WEST RAJSTHAN	**12**	**9.2**	**10.1**	**0.43**	**17.0**	**26.4**	**6.8**
EAST RAJSTHAN	**13**	**14.0**	**15.3**	**0.41**	**16.3**	**61.7**	**5.1**
WEST MADHYA PRA.	**14**	**15.5**	**16.3**	**0.35**	**9.6**	**90.1**	**8.1**
EAST MADHYA PRA.	15	19.4	19.6	0.25	2.0	112.0	7.8
GUJRAT	16	26.9	27.4	0.22	3.6	84.2	3.0
SAURASHTRA & KUTCH	17	19.3	19.5	0.22	2.0	43.0	3.8
KONKAN AND GOA	**18**	**38.7**	**44.3**	**0.50**	**23.7**	**247.0**	**1.2**
MADHYA MAHARASHTRA	**19**	**9.8**	**10.5**	**0.39**	**12.9**	**57.6**	**4.0**
MARATHWADA	**20**	**15.7**	**18.8**	**0.56**	**30.3**	**70.1**	**2.2**
VIDARBHA	**21**	**16.5**	**17.9**	**0.38**	**15.0**	**91.6**	**3.4**
COASTAL ANDHRA PRA	22	12.0	12.2	0.26	3.3	53.2	3.2
TELANGANA	**23**	**15.4**	**17.4**	**0.49**	**21.7**	**73.8**	**4.0**
RAYALASEEMA	24	11.6	11.8	0.20	3.4	43.8	2.4
TAMIL NADU	25	7.4	7.3	0.07	-2.2	31.1	4.5
COASTAL KARNATAKA	**26**	**48.2**	**50.3**	**0.33**	**8.2**	**294.0**	**0.6**
N. INT. KARNATAKA	**27**	**10.5**	**11.7**	**0.46**	**19.5**	**61.0**	**2.8**
S. INT. KARNATAKA	28	9.8	9.9	0.18	1.8	51.0	3.2
KERALA	29	34.7	34.6	0.17	-0.6	188.0	1.3
ALL INDIA	**30**	**7.8**	**8.5**	**0.41**	**16.4**	**84.1**	**100.0**

Table 1: The cross validated forecast skill of JJAS rainfall in each subdivision based on a least squares prediction using the optimal subset of predictors in the period 1951-2002. Specifically, the table gives the root mean square error (rmse), correlation skill (corr), and explained variance (EV; defined in (5)) of the statistical forecast, the standard deviation (stdv) and mean (mean) of JJAS rainfall in each subdiv, and the fractional area coverage (weight) of each subdivision used to compute areal average rainfall in India. Subdivisions with statistically significant correlation skill (corr ≥ 0.32) are in bold.

subdiv	npred	dtend	nino34mam	naojf	naoam	qbojfm	wpacmam	eindmam	arabmam	teurodjf	tindiamam
NORTH ASSAM	2						X		X		
SOUTH ASSAM	3			X			X			X	
SUB-HIMA. W. BENGAL	2		X					X			
GANGETIC W. BENGAL	3			X	X						X
ORISSA	0										
BIHAR PLATEAU	2				X						X
BIHAR PLAINS	1										X
EAST UTTAR PRADESH	5	X	X					X	X	X	
WEST U.P. PLAINS	2	X			X						
HARYANA	6	X	X	X	X				X	X	
PUNJAB	6	X	X	X	X					X	X
WEST RAJSTHAN	3	X								X	X
EAST RAJSTHAN	3	X		X						X	
WEST MADHYA PRA.	5	X					X	X	X	X	
EAST MADHYA PRA.	3			X						X	X
GUJRAT	1	X									
SAURASHTRA & KUTCH	2	X			X						
KONKAN AND GOA	3	X					X		X		
MADHYA MAHARASHTRA	5	X	X				X		X		X
MARATHWADA	4	X	X				X		X		
VIDARBHA	1	X									
COASTAL ANDHRA PRA	5	X	X				X	X			X
TELANGANA	4	X		X			X			X	
RAYALASEEMA	1	X									
TAMIL NADU	2	X	X								
COASTAL KARNATAKA	3	X			X					X	
N. INT. KARNATAKA	4	X	X				X	X			
S. INT. KARNATAKA	1	X									
KERALA	2				X					X	
ALL INDIA	3	X			X						X

Table 2: List of Indian subdivisions and the specific predictors, indicated by an X in the appropriate table entry, which give the optimum cross validated mean square error of JJAS rainfall in that subdivision for the period 1951-2002.